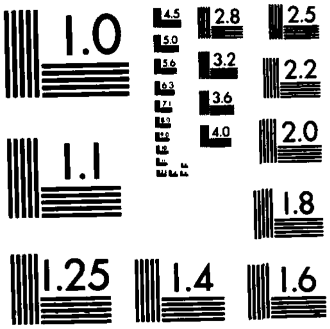
AD-A123 342 BASIC TECHNIQUES IN ENVIRONMENTAL SIMULATION(U) AIR FORCE ENVIRONMENTAL TECHNICAL APPLICATIONS CENTER SCOTT AFB IL R C WHITON ET AL. JUL 82 USAFETAC/TN-82/004 1/2

UNCLASSIFIED SBI-AD-E850 217 F/G 4/2 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

USAFETAC/TN-82/004

BASIC TECHNIQUES IN ENVIRONMENTAL SIMULATION

BY

Lt Col Roger C. Whiton

Capt Emil M. Berecek

JULY 1982

Approved For Public Release; Distribution Unlimited

UNITED STATES AIR FORCE
AIR WEATHER SERVICE (MAC)
USAF
ENVIRONMENTAL
TECHNICAL APPLICATIONS
CENTER
SCOTT AIR FORCE BASE, ILLINOIS 62225

REVIEW AND APPROVAL STATEMENT

USAFETAC/TN-82/004, Basic Techniques in Environmental Simulation, July 1982, is approved for public release. There is no objection to unlimited distribution of this document to the public at large, or by the Defense Technical Information Center (DTIC) to the National Technical Information Service (NTIS).

This technical publication has been reviewed and is approved for publication.

FOR THE COMMANDER

Patrick J. Breitling

DR. PATRICK J. BREITLING
Chief Scientist

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER USAFETAC/TN-82/004	2. GOVT ACCESSION NO. AD-A123 342	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) BASIC TECHNIQUES IN ENVIRONMENTAL SIMULATION		5. TYPE OF REPORT & PERIOD COVERED Technical Note
		6. PERFORMING ORG. REPORT NUMBER USAFETAC Project 2082
7. AUTHOR(s) Roger C. Whiton, Lt Col, USAF Emil M. Berecek, Capt, USAF		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS US Air Force Environmental Technical Applications Center/DNS Scott AFB, Illinois 62225		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS US Air Force Environmental Technical Applications Center Scott AFB, Illinois 62225		12. REPORT DATE July 1982
		13. NUMBER OF PAGES 144
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Algorithms, autocorrelation, autoregressive moving average model, Burr curve, ceiling, Cholesky reduction, climatology, computerized simulation, correlation, cumulative distribution function, curve fitting, distribution fitting, environmental simulation, fitting, forecasts, gaming, great circle distance, joint probability, marginal probability, Markov processes, mathematical (Cont'd)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Environmental simulation modeling is defined as the generation of synthetic weather observations and forecasts by use of mathematical/statistical models. Basic concepts in environmental simulation modeling are described, with emphasis on underlying statistical fundamentals, stochastic processes, and Markov processes. Four principal environmental simulation models and their application are described in detail. The treatment begins with the single-variable, single-station model, V1S1, and is extended to the two-variable, single-station model, V2S1. The (Cont'd)

19. KEY WORDS (Cont'd): models, meteorology, models, multivariate normal, MULTRI (Multivariate Normal Triangular Matrix Model), normalization, partial autocorrelation, probability, probability density function, probability distribution, probability distribution fitting, probability function, random normal numbers, random normal vectors, random processes, random variable, reverse Weibull distribution, root mean square difference, sawtooth wave, simulation, sky cover, serial correlation, spatial correlation, statistics, stochastic modeling, stochastic processes, temporal correlation, tetrachoric correlation, uncertainty, visibility, V1S1 (Single-variable, Single-station Model), V2S1 (Two-variable, Single-station Model), war gaming, weather, Weibull distribution, 2DFLD (Two-dimensional Field Simulation Model).

20. ABSTRACT (Cont'd): multivariate triangular matrix model, MULTRI, is then discussed; that model is capable of generating vectors of N correlated variables. A case study is presented showing the application of MULTRI to modeling moint sky cover distributions at station pairs or at a single station for N lag times. The most complex model in the series of four is the two-dimensional field simulation model, 2DFLD, capable of producing spatially correlated, synthetic, two-dimensional fields or networks of variables. Statistical methods used in developing environmental simulation models are described, with particular emphasis placed on how to fit probability distribution functions to weather variables.

PREFACE

This report is a description of environmental simulation techniques developed for USAFETAC Project 1960, 2082, 2339, and 2357. The report provides a basic description of current USAFETAC modeling capabilities and serves as a tutorial for practitioners and users of environmental simulation modeling.

We gratefully acknowledge the numerous suggestions and contributions of Major Albert Boehm, USAFETAC/DNP.

CONTENTS

PAGE

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 General

On 26 February 1979 the United States Air Force Technical Applications Center (USAFETAC) was designated as the focal point for providing Air Weather Service's (AWS) environmental simulation support. This new mission for USAFETAC resulted from growing environmental simulation requirements within AWS. It was felt that if simulation support were handled through a centralized facility, the techniques developed for one customer could be applied to others. Centralized support would make the simulation expertise available to all AWS personnel and not tie it to the life cycle of individual projects.

This technical note describes some of USAFETAC's initial modeling capabilities and should provide the reader with some fundamental statistical background that will be needed for more advanced modeling problems.

1.2 Environmental Models

For the purpose of this technical note the term environmental refers only to meteorological applications of modeling. Other disciplines such as geophysics, hydrology, and engineering have done quite a bit of work in modeling their own spheres of interest and by rights should be included under a term dealing with man's environment. Such, however, is not the scope of this report.

There are two types of environmental models. One type of model is based on a mathematical representation of the dynamics of a real life system. These models are dynamical initial-boundary value problems. Once the initial conditions of the system have been determined, the state of the system at any future point is given by the analytical or numerical solution of a set of differential equations. These equations are based on physical laws of nature such as the laws of motion. Examples of these types of models are the NOAA National Weather Service's Limited Area Fine Mesh (LFM) Model and the Air Force Global Weather Central's Boundary Layer Model (BLM). For some problems, finding analytical or numerical solutions to dynamical models is too arduous, and one must consider a second type of modeling, namely simulation.

Environmental simulations apply the theory of mathematical statistics to mirror the processes and interrelationships of a real life system. While these models may do very well in producing certain desired statistics such as means, standard deviations, and correlations, they might violate physical laws.

Environmental simulations range from deterministic models (i.e., ones in which, given the current state of the system, the future state is uniquely defined) to purely stochastic models (i.e., ones in which the system behavior is inherently uncertain or random). Most environmental simulation models, however, are a mixture of the two. While the current state of the system might weigh heavily on the results, because of the uncertainty of weather events, the outcomes are not always the same given the same input data.

1.3 Environmental Simulation

Weapons systems effectiveness studies, design trade-off analyses, combat tactics, strategy and doctrine development simulations, war games, and other similar activities often need some type of weather input. This weather input is to test whether the resulting weapons systems or war plan will work correctly in real combat weather. At least two approaches can be used to provide the weather input to these weapons and warfare design studies: (1) the historical weather record (climatology) could be used directly to provide sequences, means, standard deviations, joint probabilities, and the like for stations and areas of interest; or (2) the relevant weather variables could be mathematically modeled in space and time and then this model used to infer the needed quantities. The model could even be used to generate desired time series of the critical weather variables at selected stations or over specified grid systems.

The latter approach is variously referred to as modeled climatology, synthetic meteorology, or environmental simulation. Whatever the term applied, the technique involves using mathematical and probabilistic models to achieve a selectively realistic synthesis of the environment in order to describe or analyze the environment or the effects of the environment on a system or operation.

How is this done? Typically, a stochastic (i.e., random process) model is found that can produce synthetic weather "data" realistic enough to meet the user's needs. Such "data" might, for example, be required to have the same means, variability, and cross-correlations as are found in the "real" data. Then the model is coded as a computer program or subprogram. Next, the model is "fitted" to the weather by consulting the historical weather record to find regression coefficients, correlations, probability distributions, and other model parameters. Then the model is tested for validity against an independent sample of historical weather data. Finally, the model is used to generate a long series of synthetic data, and the data are analyzed to answer the planning or design questions at hand. Wherever possible, the environmental simulation model is coded as a subroutine or subprogram within the user's larger model. The weather routine then generates and delivers synthetic weather information -- either observations or forecasts -- whenever "called" by the larger model.

Mathematical/statistical models such as the one hinted at above are stochastic rather than deterministic because they treat the weather as a partly random (stochastic) process. The approach described above is numerical rather than analytical because it makes use of approximative or iterative methods to converge toward a solution. Moreover, within the class of numerical mathematical models, this approach would be referred to as a Monte Carlo technique because it involves the use of statistical sampling methods (especially drawing random numbers) to obtain a statistical parameter or other probabilistic solution to a physical/mathematical problem. One need not always resort to Monte Carlo techniques in simulation modeling. Although the Monte Carlo approach is attractively simple and flexible, it is on the other hand computationally expensive and produces only approximate solutions. Some problems are amenable to analytical solution. In this approach, needed equations are derived from the theorems of mathematical statistics. These equations are then simply evaluated to produce the statistical parameters required to answer the planning or design questions at hand. There is no need to generate a long series of synthetic weather observations or forecasts if one is using an analytical environmental simulation model. In practice, both numerical/Monte Carlo methods and the analytical method are used in environmental simulation, as they are in simulations of all kinds.

The question may arise, why use simulation? USAFETAC has over 80,000 magnetic tapes of worldwide weather data. Why use valuable resources developing simulation models if raw data are so abundantly available? The reasons are manifold.

1.4 Environmental Simulation vs. Direct Use of Historical Weather Data

The main reason for using simulation models rather than using historical data directly is that most users of weather information are designing new weapons systems or planning for future wars, not conducting post-mortem analyses of old ones! From the point of view of statistical sampling theory, the historical weather record is almost always very short. It probably does not contain all the patterns of weather likely to occur in the future. The worst weather on record is not the worst possible weather. Any historical record is but one realization of a stochastic time series, and future realizations will resemble that historical record only in a statistical sense, even if the underlying probability distributions and cross correlations do not change.

In using an actual historical weather sequence in planning or design, one runs the risk that the sequence chosen may be too mild, too severe, too bland, too erratic, or in some other way unrepresentative of the future. Moreover, if a particular year or month is chosen as "typical weather," it is often the only weather that is run in support of the analysis being conducted. The "typical weather" file can be used and reused repeatedly in planning or design, and the plan or design can be tuned so closely to that particular weather sequence that the plan or design becomes virtually inapplicable except to that particular year or month in history. The danger in using historical weather data directly is particularly great when the data will be used in a war game. If the weather always turns up bad on 15 December and always improves on 20 December, the war gamers will soon begin to notice this and take advantage of the unnatural lack of "weather surprises" in their combat development planning. Using canned weather in war games can lead to the development of "optimum" tactics and force mixes that will not withstand the test of actual employment in real, future weather.

Using an environmental simulation model rather than the historical weather record directly helps circumvent problems such as these. The model will produce synthetic weather "data" (observations, forecasts, or both) on call. The user controls the length of the time series generated. The synthetic weather possess a variability not unlike that of real weather data, in the sense that the weather for one 15 December will not be the same as that for all other 15 Decembers. This quasi-natural variability permits running the model repeatedly to acquire risk statistics needed by the designer or planner.

Still another reason for using simulated or synthetic weather instead of historical weather data, is that environmental support requests are becoming increasingly complex. For many applied climatological problems, such as probabilities over an area or probabilities along the line, direct use of the data base may not produce needed answers. USAFETAC is often asked for data or answers at arbitrary locations or at grid points for which no weather data exists at all.

At other times, the customer's request is for a mission-dependent or system-dependent weather effects parameter such as DCFLOS, the probability of a cloud-free line-of-sight between two moving points A and B for a specified duration of time. The dynamic cloud-free line-of-sight (DCFLOS) probability depends just as much on the movement of A and B and on the lock-on duration as it depends on the cloud cover. Evaluating these system-dependent weather effects without using some sort of simulation is usually impossible.

Other impelling reasons for using an environmental simulation model rather than the "real" data are the inaccuracies and inadequacies in the historical weather data base and the simple convenience of having the weather generated by a small, fast computer subroutine rather than by reading and rereading an extensive tape- or disk-based data set.

1.5 Environmental Models Developed from Real Data

It should be emphasized that environmental simulation modeling is not done in the dark, without regard to the historical weather record. USAFETAC makes heavy use of the historical weather record to produce parameters such as probability distributions and correlations needed by the mathematical model. Actual weather data is used extensively to test and verify the simulation model once it is built. In this sense, the historical weather record is used indirectly for the studies and games because the model is developed and tested using climatology.

Now that the reasons USAFETAC is involved in environmental simulation have been discussed, let us define some basic terminology.

1.6 Glossary of Terms in Environmental Simulation and Related Areas

Environmental Simulation: A selectively realistic synthesis of aerospace behavior consistent in space and time, achieved by the use of techniques -- often involving mathematical and probabilistic models -- with which to describe or analyze the environment or the effects of the environment on a system.

Gaming: A gaming exercise employs human beings acting as themselves or playing simulated roles in an environment that is either actual or simulated. The players may be experimental subjects or participants in an exercise being run for teaching, operational training, planning, or other purposes.

Model: A model is a representation, description, or imitation of a system or process (e.g., the atmosphere) in another medium (e.g., a computer). A model is a generalization of a more complex reality usually involving simplifying assumptions in order to produce understandable solutions. A good model is constructed so as to produce realistic behavior critical to the problem at hand while preserving the essential properties of the system being simulated.

Simulation: A simulation is an analytical or numerical technique involving the use of mathematical and logical models to represent and study the character and behavior of real-world or hypothetical events, processes, or systems, over extended periods of time. Simulation enables a real system or process to be studied, analyzed, and understood by means of a model. All simulations involve models, but not all models are simulators. Simulation is usually done for such purposes as training, experimentation, evaluation, and finally, to draw conclusions about the system or process being simulated. Simulation provides the means for gaining experience and for making and correcting errors without incurring the costs or risks of actual application. It offers opportunities to test theories and proposed modifications in systems or processes; to study organizations and structures; to probe past, present and future events; and hypothetically to utilize forces that are difficult or impracticable to mobilize. Simulation, therefore, is of value both as an educational device and as a means of discovering improved methods. The distinction between games and simulations is sometimes confusing. Games use a simulated environment or simulated roles for the players, or both. In general, all games are simulations, but not all simulations are games. Computer simulations that model conflict or cooperation (such as completely computerized battle models) are usefully considered as games. Possibly, so are some logistic or resource allocation models where the single (automated or live player) team may be regarded as struggling against a statistical or strategic opponent called "Nature," although here one enters the territory of decision theory. The borderline is not hard and fast; however, it is probably not useful to treat a straight industrial production scheduling machine simulation as a game.

War Gaming: A war game is a simulation of a military operation involving two or more opposing forces and using rules, data, and procedures designed to depict an actual or assumed real-life situation. It is primarily a technique used to study problems of military planning, organization, tactics, and strategy. A war game can be accomplished manually, can be computer-assisted, or be wholly computerized. Manual games are played using symbols, pins, or pieces to represent forces, weapons, and targets on maps, mapboards, and terrain models. A computer-assisted game is a manual game using computerized models that free the control group from many repetitive, time-consuming bookkeeping computations. Computerized war games are based on predetermined procedures and rules, and all simulation of conflict is done by the computer in accordance with the detailed instructions contained in the computer program. The primary advantage of computer gaming is that the same situation can be simulated many times under differing conditions, in order to observe the variability of results.

1.7 A Note to Users of Environmental Simulation Models

1.7.1 Project Success--A Shared Responsibility. Users share with model developers very real responsibilities for the success or failure of simulation projects of all kinds, including environmental simulation efforts. The potential simulation user's concept of how weather should "play" in a particular study, the user's views as to what environmental simulation modeling can and cannot do, and his opinions on how best to use weather simulation dominate the scene during the critical early stages of problem definition. Often these early conceptions regarding what needs to be done and how to do it persist--for better or for worse--throughout the entire lifetime of the project. These "preliminary ideas" quickly set up like concrete. If the ideas are well thought out, they can serve as a substantial foundation for the project as well as a true template for the project's future growth. If, on the other hand, the user's ideas are incorrect, unrealistic, or out-of-date, they can imprison a project, stunt its growth, and seriously impair its chances of success.

Because the user's role is of such dominating importance during the early stages of a project, some attention is given here to developing a common understanding among the community of actual and potential users regarding such questions as:

- What can environmental simulation modeling do and what can't it do?

- Under what circumstances would the user be better advised to use the historical weather record directly?

- In writing a requirement for an environmental simulation model, what can and cannot be asked for? In what terms does one specify the requirements?

- How do requirements for model design, performance, format, interfacing, and documentation affect the cost of the project in time and money?

1.7.2 *What Can and Cannot Be Done in Environmental Simulation Modeling*? Stating what can and cannot be done in any scientific or technical field is a hazardous undertaking; for the state of science, mathematics, and technology is subject to change. All that can be done with any confidence of being right is to summarize the state of the science *today*.

Projects whose requirements extend beyond the limits of today's scientific, mathematical, or statistical techniques are said to require at least *technique development* and quite possibly *applied* or even *basic research* before they can be satisfied. While projects requiring such advancements in the state of the science can be done (basic or applied research must, of course, be accomplished by the Air Force Systems Command). Such projects will normally incur a much greater risk of failure and, even if successful, will ordinarily be much more costly and time-consuming to complete than projects that require little if any advancement in knowledge. Today's state of the science in environmental simulation modeling is described below.

It is today possible to generate by mathematical/statistical models time series of synthetic weather observations and forecasts at a single point, over an irregularly spaced network of points, or over a regularly spaced, two-dimensional grid of points, provided that sufficient historical weather information exists with which to estimate the statistical character of observed and forecast weather at the locations involved. The models can generate univariate or multivariate synthetic data. A long run of synthetic weather observations produced from such an environmental simulation model will, in terms of certain statistical measures, be indistinguishable from a comparable run extracted from the historical weather record. The statistical measures "preserved"[1] by the environmental simulation models described in this technical note include:

[1] An environmental simulation model is said to "preserve" a statistic such as a probability distribution or a correlation if that statistic, computed from synthetic data generated by a sufficiently long run of the model, is not significantly different from the same statistic computed from a sufficiently long period of the historical weather record.

- Unconditional cumulative distribution functions of the variables being simulated

- Serial correlation of each variable over the index parameter t (usually representing time) of the simulation

- Cross correlation between variables when the simulation model is of multivariate design

- Spatial correlation in two dimensions only

- Skill of weather forecasts

1.7.3 Requirements Drive the Solution: Models or Data? Under some circumstances--for example, weather information used in exercises and training simulations--weather observations and forecasts are required to have exceptional synoptic "realism." Often under these circumstances a complete meteorological "scene" is required, involving organized, moving, evolving cyclone and frontal systems with horizontally and vertically consistent dynamical and thermodynamical fields and supporting three-dimensional cloudiness patterns. Such requirements are stated because in a training exercise, actual meteorological displays are prepared, much like those in weather stations. Weather briefings are given and simulated forecasts made from these displays. The whole package has to look "realistic" from the user's point of view; otherwise, the realism or even the credibility of the exercise or training simulation is to some extent compromised.

No statistical simulation model has yet been developed capable of generating multivariate, multicorrelated time series of three-dimensional weather "scenes." A user whose legitimate requirements call for meteorological realism of this degree must employ the historical weather record directly. In doing so, the user is subject to all the difficulties and limitations discussed above, associated with direct use of the historical weather record.

Far more often than not, however, the study, analysis, simulation, or game being conducted has no need for meteorological "realism" of this degree. Consider, for example, a simple Monte Carlo air reconnaissance simulation that generates and scores individual reconnaissance sorties involving takeoff from location $A(t_0)$, air photography at locations $B(t_1)$, $C(t_2)$, and $D(t_3)$ and return for landing at location $A(t_4)$, with alternate at $E(t_5)$, where t is the time parameter. In this model, weather is used simply to "score" the mission's success; it is not used for mission planning. Weather impacts are shown in Table 1.

Table 1. Weather Impacts in a Hypothetical Air Reconnaissance Simulation.

Location (Time)	Criteria for Success or Probability of Success
A(t_0)	Takeoff Requirement: Ceiling/Visibility ≥ 200 ft/½ mile

B(t_1)
C(t_2)
D(t_3)
	Reconnaissance Photography Requirement:
	$\Pr\{\text{Success}\mid H_C\} = (0.95/3500)(H_C-1500)$ where
	$0.0 < \Pr \leq 0.95$ and H_C=ceiling height (ft)

A(t_4)	Landing Requirement: Ceiling/Visibility ≥ 200 ft/½ mile
E(t_5)	Alternate Requirement: Ceiling/Visibility ≥ 200 ft/½ mile

The mission will launch if the takeoff requirement is satisfied, will attempt to film all three targets regardless of weather but with a probability of success that rises linearly from zero with ceilings of 1500 ft or less to 0.95 with ceilings of 5000 ft or more, and will land at airfield A, provided the landing requirement is satisfied there at time t_4. Otherwise, the aircraft will proceed to alternate airfield E at time t_5 and will land there if the weather is good or experience a 65-percent chance of abort with loss of film if the weather is bad.

To support this simple reconnaissance simulation, it is necessary only to supply ceiling "observations" (historical or synthetic) at these five locations at the times indicated, as well as visibility observations at A(t_0), A(t_4), and E(t_5). The correlation between the weather at one location and that at another decreases with increasing distance between the points. Therefore, the weather at points such as A and E, and such as points B, C, and D cannot be treated as independent in space. Some sort of distance-dependent spatial correlation must be built into the weather information, synthetic or historical, that is supplied to the reconnaissance simulation. Since the mission extends over a duration of time, either (t_4 - t_0) or (t_5 - t_0), consideration must be given to the time-continuity of weather supplied to the reconnaissance simulation. A statistical measure of this continuity in time is the so-called serial correlation. The serial correlation of the weather information delivered to the reconnaissance simulation must be patterned after that observed in nature. At certain locations, namely A and E, the visibility as well as the ceiling must be supplied. Data studies show that the ceiling and visibility are positively correlated. Hence, ceiling, and visibility pairs supplied to the reconnaissance simulation must, in the long run, demonstrate this so-called cross correlation between variables. Finally, the ceiling and visibility information delivered to the reconnaissance simulation should not in the long run violate the probability distributions of the ceiling and the visibility for the time and place of the simulation. In other words, the sample ceiling and visibility information supplied to the reconnaissance simulator must be drawn from the same populations that the longer-term historical weather record was drawn from.

In summary, the weather information supplied to this hypothesized air reconnaissance simulator must exhibit appropriate probability distributions, spatial correlation, serial correlation, and cross correlation. These are sufficient requirements to be imposed on the weather used for this application. Nothing in the planned use of weather by this hypothetical reconnaissance model suggests a need for cyclones, fronts, spiral band cloud patterns, and the like. As long as the weather information supplied to the reconnaissance simulation has "realistic" probability distributions, spatial correlation, serial correlation, and cross correlation, it should be sufficient to meet the need.

In this case, an environmental simulation model could be used to generate synthetic ceiling and visibility data with appropriate probability distributions and correlations. If additional, and in this case superfluous, requirements for synoptic realism were to be imposed, models could not be used, and historical weather data--with all their limitations--would then have to be resorted to. The effect of adding superfluous requirements would in this case be to force a suboptimal solution. In general, potential users of environmental simulation models should study in detail how their applications model uses (or proposes to use) weather information and then state their requirement as conservatively as possible, expressing the requirement in terms of the statistical measures that an environmental simulation model must "preserve" (see footnote 1 above).

1.7.4 Model Decisions and the Need for Weather Observations and Forecasts. Applications models such as weapons systems effectiveness simulations and combat evaluation models use weather information to make decisions that emulate those made in real time by human decision makers (such as battle staffs and individual aircraft commanders) and the "forces of chance and nature" (such as whether a particular reconnaissance target is photographed, given the weather). Such models, even if they "play" only one side of the combat, generally have to assess the consequences of decisions made by the side whose actions are being "played." This is referred to as mission assessment or "scoring" and represents the most common use of weather in military studies and analyses. In scoring, the model makes a decision, based on weather and other factors, as to whether, for example, a given reconnaissance target is successfully "shot" by aerial photography. For scoring decisions impacted by weather, applications models need the value of mission-critical weather variables at the time the mission is executed.

The concept of "scoring" missions based on weather can be extended to include other impacts of weather on mission execution, such as enroute winds affecting a simulated airlift mission's flight time or protracted rainfall slowing the rate of advance of an armored column. Scoring decisions made by military applications models tend to emulate the impersonal aspects involved in the course of military events, such as assessing the success of missions, governing the timing of an advance, or determining other partly probabilistic outcomes. *In real life*, there is no need for scoring decisions. They are made for us by the "forces of chance and nature" or by action of opposing forces. But in a simulator, these chance outcomes, natural impacts and effects of enemy action must be included in the simulator, or they simply will not occur.

In the simplest combat simulations or weapons system effectiveness studies, "scoring" decisions are the only ones made using the weather. In other models, an attempt is made to emulate selected aspects of the human decision-making process as applied in combat. From a meteorological perspective, human decisions can be classified as either (1) short-range execution decisions based on the observed present weather, or (2) longer range planning decisions based on future weather. To model the influence of weather on the spectrum of human decisions from planning to execution requires that the applications model consider not only weather observations but also weather forecasts. Decision points have to be built into the applications model so as to call for and use weather observations and forecasts much as they are used in the actual or proposed system being modeled.

In practice, few of today's weapons systems effectiveness models, combat development simulations, and other applications models consider even the observed weather, and almost none of them (except those built by meteorologists themselves[2]) use forecast weather. This situation is changing, however. The U.S. Air Force Air War College operates a combat model which, in the 1970s, was modified to accept statistically generated weather forecasts for input to decision making. In the late 1970s, a statistical model that generates synthetic weather observations and forecasts was added by USAFETAC to the Military Airlift Command's *M-14 airlift system simulation. In 1981, USAFETAC designed* a statistical, two-dimensional field simulation model to generate cloud forecast fields for input to system planning and optimization models. Environmental simulation modeling efforts such as these have received increasing attention since 1979 in such media as the Air Weather Service Operations Digest (see January-February 1981 issue) and the 2nd Weather Squadron Technical Activities Summary (see July 1980 issue), as well as in the Military Operations Research Symposia (B-4 Working Group presentation at 46th MORS, December 1980, and general session presentation at 48th MORS, December 1981), and in American Meteorological Society conferences (6th AMS Conference on Probability and Statistics in Atmospheric Sciences, October 1979). These efforts at communicating what has been done in environmental simulation modeling should have the effect of showing the military modeler what is possible and increasing his interest in factoring weather effects into his models.

[2] One of the earliest applications models to include forecasts was in fact built by meteorologists to support the Weather-85 Mission Analysis of the Air Weather Service. See Huschke, R. E., and R. R. Rapp (1970): Weather-Service Contribution to STRICOM Operations--A Survey, A Model and Results: Final Report on Phase I of the Rand Corporation Contribution to the Air Weather Service Mission Analysis, R-542-PR, The Rand Corporation, 58 pp.

1.7.5 Stating Requirements for Environmental Simulation Models. Given that a need exists for the sort of weather information that could perhaps be provided by an environmental simulation model, one of the first things that must be done is to express that need in the form of a requirement.

In some cases, the user of weather information will not care whether that information comes from a model or directly from the historical weather record, just as long as the information is "good enough" to meet his needs. In other cases, an environmental simulation model will be explicitly called for. Under both circumstances, the user's job is to state his requirements in terms of the variables to be provided, whether forecasts, observations, or both are required, the time and space dimensionality[3] of the information needed, the statistics[4] to be preserved, and the accuracy required, expressed in terms of some standard relevant to the user's problem.

1.7.6 Technique Development vs. Software Development. After a requirement for weather information has been stated and it is determined that an environmental simulation model is the most effective way to proceed, the user should specify whether he needs simply a tested proven technique or finished software. In the former case, the product delivered is generally a complete description of the technique and an analysis of its performance, accompanied by a courtesy copy of the exploratory software developed to test the model, for the latter case, a software development phase is added to the project. In that phase, technique development software is converted to fully qualified, fully maintainable, fully documented software in strict accord with the Air Force 300-series software management directives. The user should be aware that although the final product is much more polished in the latter case, considerable time is added to the project completion estimate in order to comply with the software management requirements.

1.7.7 Operational Environment, Interfaces, and Constraints. In stating requirements for environmental simulation models, it is useful to specify (1) the operational environment of the model, e.g., whether the model is to stand alone or is to serve as a module within a user's larger model and the computer and operating system on which the model is to run; (2) the interface between the environmental simulation model and the user models, applications or studies the model is to serve, i.e., inputs and outputs required and whether the environmental simulation model will reside within and be called by the larger user applications model; and (3) constraints within which the environmental simulation model must exist, i.e.,

[3] For example, single-station, two or three spatial dimensions, irregularly spaced network, regular grid, time continuity, required or not, etc.

[4] Unconditional probability distributions, conditional or joint probabilities, correlations in time and space, cross correlations between variables, means, standard deviations, etc.

computer programming language requirements, computer program data structure requirements, computer program size, and speed constraints, etc.

1.7.8 Effectiveness Evaluation, Value Analysis, and Feedback. Effective feedback provided by those who state requirements for and use models to those who develop them is almost surely the best means of improving the whole model development process. Users receiving environmental simulation models should test them to determine whether they meet requirements. Usually the developer facilitates such testing by leaving in the delivered model certain test modules that measure key aspects of the model's performance. As a first step, the user can repeat and verify those tests in a stand-alone environment, in which the environmental simulation model is not yet interfaced or integrated with the larger user model it is to serve. More important from the user's perspective, however, is his unique ability to test the environmental simulation model in an integrated environment within the larger user application. The developer usually cannot perform these invaluable integrated tests because he does not have access to or familiarity with the larger model.

Results from stand-alone and integrated testing performed by the user should be communicated quickly to the developer, especially when those tests indicate changes must be made to the model. The time to make these changes is right away, not 6 months after delivery. By then the developer has gone on to other work and has lost his familiarity with the model. Ordinarily, the developer will provide a 90-day warranty on models and software. During that 90-day period, the developer is liable for all necessary changes in the model or its supporting computer software. After the 90-day warranty expires, the user, not the developer, is responsible for all changes. The user should therefore finish all stand-alone and integrated testing before expiration of the 90-day warranty.

Developing a simulation model of any sort is an expensive undertaking, requiring a great many manhours and computer hours for development and testing. Under these circumstances, it is helpful to receive from the user information describing the benefits derived from use of the environmental simulation model. In some cases, adding simulated weather to a study, analysis, or plan improves decisions quantifiably -- for example, by causing abandonment of a weapons systems design which, if carried through, would have been an expensive failure, or showing how by intelligent use of weather information an airlift activity can increase the tonnage hauled. Information of this sort is useful in establishing the cost effectiveness of environmental simulation modeling and in justifying its continued use.

1.8 Basic Environmental Simulation Concepts, Techniques, and Procedures

The remainder of this technical note consists of a description of key concepts in statistics and simulation, followed by a description of USAFETAC's most basic and most generally useful environmental simulation models:

- Single-station, Single-variable Ornstein-Uhlenbeck Model (V1S1)
  (Chapter 3, Basic Single-station Models)

- Single-station, Two-variable Ornstein-Uhlenbeck Model (V2S1)
  (Chapter 3, Basic Single-station Models)

- Multivariate Triangular Matrix Model (MULTRI)
  (Chapter 4, Multi-parameter/Multi-station Models)
  (Chapter 5, Modeling Joint Sky Cover Distributions)

- Two-dimensional Field Simulation Model (2DFLD)
  (Chapter 6, A Model for the Simulation of Gridded Fields)

Chapter 2

BASIC CONCEPTS IN ENVIRONMENTAL SIMULATION MODELING

2.1 Uncertainty in Science

Although many scientific problems are solved deterministically, as if the scientist could predict the outcomes of his experiments with certainty, nevertheless the "real world" of science is based on and has come to terms with uncertainty or indeterminancy.

For many real-world problems in science, solutions cannot be stated deterministically, or, what is more often the case, the deterministic solution is only an approximation to the complete solution. In many cases, deterministic solutions represent the expected value of the true solution or even worse, just one of a spectrum of possible values constituting the true solution.

Circumstances such as these prevail widely in studies of distinctly *random processes*, i.e., processes whose outcomes are uncertain, or processes having a number of possible outcomes, each with its own *probability*. Examples arise from the study of molecular motion, atomic decay, and other physical processes whose character is inherently *statistical* or random.

Uncertainty or indeterminacy in science is not restricted to the small world of atoms and molecules but rather extends itself to much larger phenomena such as atmospheric turbulence, which must be treated probabilistically, and even onward to the bulk parameters of the earth's atmosphere at large, such as temperature, density, and pressure. In the final analysis, the definitions of these bulk variables are inherently statistical, being based on the fleeting presence and motion of molecules in the sampling volume.

If one could station oneself inconspicuously as an ever-so-small floor walker in such a sampling volume, then one would see at one time a few molecules in the volume and at other times many; some moving slowly, others moving fast; at times colliding, and at other times not. Lucretius said it best more than 50 years before the birth of Christ

> ... Nor did they bargain sooth to say what motions each should assume but because many in number and shifting about in many ways throughout the universe, they are driven and tormented by blows during infinite past. After trying motions and unions of every kind, at length they fall into arrangements such as those out of which this our sum of things has been formed....
>
> -- *De Rerum Naturae*

Lucretius' final point is his most important one. When individually unpredictable events such as molecular collisions, turbulent eddy motions, or "fair" coin tosses are repeated at length, there usually emerges some form of regularity or appears some aggregate result, such as temperature, a cascade of energy, or a probability of 1/2.

The idea of <u>probability</u> is at the heart of the concept of a random process, because there is uncertainty in the outcome of such a process, and each outcome has an associated probability.

There is a physical/mathematical link between deterministic problems and random processes. In many cases, it is possible to write down deterministic problems in terms of partial differential equations that yield a distribution function for the probabilities associated with the process. As it turns out, the partial differential equation for the distribution function is the same as the partial differential equation that would be given in a deterministic statement of the phenomenon (Lin and Segel, 1974).

In fact, physical processes that differ from each other greatly in detail -- such as Brownian motion, heat conduction and diffusion of one gas through another -- are all described in the limit or in bulk by the <u>same</u> partial differential equation.

It is frequently the case in science that a given problem can be stated either deterministically or probabilistically, depending on the phenomena being studied or the sort of analysis being conducted. If the element of uncertainty or incompleteness of information is high, with many other partially known factors contributing to the outcome, then the process might better be considered random rather than deterministic. Similarly, if a slight change in these contributing factors or initial conditions could potentially lead to a large change in the final outcome, then this <u>sensitivity</u> of the problem also argues for a random process treatment.

2.2 <u>Uncertainty in Meteorology</u>

2.2.1 <u>Nonlinear Interactions Make the Atmosphere a Continuum</u>. The atmosphere is a continuum in which every scale of motion affects every other scale through the nonlinearity of the governing equations of motion. Every flap of a gull's wing anywhere on the planet must affect, however weakly, the motion of every molecule of air in our atmosphere. Small eddies and turbulent flows provide an important kinetic energy dissipation mechanism to the large-scale flow, and without this sink of energy, the larger scale would necessarily behave differently than it does. Similarly, the latent heat released by mesoscale convective processes acts as an important energy source for the larger scale flow and -- at least in the aggregate -- affects that larger flow. The spectra of time and space scales for atmospheric phenomena are continuous: motions and phenomena exist at all scales, and each scale interacts with the others in complicated ways not fully understood.

2.2.2 The Predictability of the Atmosphere is Bounded. Modern studies of the predictability of atmospheric motion have shown a sensitivity problem with the governing equations of motion. Flows starting from only slightly different initial conditions can quickly evolve to radically different final states. The studies of Lorentz have shown a limit to atmospheric predictability of about 2 weeks because of this sensitivity of the governing equations to minor initialization differences. The field of dynamic meteorology is thus confronted with the need to know how predictions of the atmospheric flow will be affected by slight changes in initial values, boundary values, and simplifications used in formulating the prediction systems themselves. In a practical sense, the predictability argument is related to the way in which the initial state of the atmosphere is observed and reported. For the most part, weather observations are taken at the synoptic scale. Phenomena whose characteristic size is smaller or whose lifetime is shorter than this scale are imperfectly described. These imperfections in the description of the initial state may take a long time to affect the flow, but eventually they will become important. As a result, weather predictions based on imperfect initial observations or simplified physical equations will eventually fail.

2.2.3 Atmosphere Not in Thermodynamic Equilibrium. If the atmosphere were in thermodynamic equilibrium, the air over the whole planet from its surface to the top of the atmosphere would be as still as on the sultriest day of summer. Nor would a single drop of life-giving rain fall anywhere on earth.

It is the nonequilibrium conditions in meteorology that cause the weather, namely the thermodynamically unbalanced system, the ageostrophic wind, the nonhydrostatically balanced, vertically accelerated motion field. Yet the meteorologist's weather forecasting models bring out these nonequilibrium states poorly, if at all. Even the analysis models, which purport to describe the observed or initial state of the atmosphere, work hard to smooth out the nonequilibrium features, as these features destablize the forecasting models.

All of this contributes to an underlying uncertainty about not only the future state of the atmosphere but also its present state. This uncertainty is most apparent in the weather variables of critical operational interest, such as sky cover, ceiling, and visibility.

2.2.4 Variables of Most Interest in Applied Military Climatology Are Subject to Important Meso- and Microscale Influences. The variables of greatest interest in applied military climatology -- such as ceiling, visibility, sky cover, boundary layer winds, and precipitation -- are subject to important, nonequilibrium meso-

and microscale influences that are beneath the resolution of today's weather observing, analysis, and forecasting systems.

2.2.5 The Climate May Be Changing. Paleoclimatological records show that the earth's climate has been subject to significantly large changes in the past. There is no reason to believe that the climate will not change in the future, if indeed it is not already changing. Nevertheless, the usual assumption in meteorological and climatological modeling is that the climate does not change. The science of climate change and climate prediction remains in its infancy among meteorologists. There is very little understanding of how and why the climate has changed in the past and almost no ability to predict when such changes will occur in the future and how great a change is to be expected.

2.2.6 Consequences. The consequence of these circumstances is that, for many purposes, weather is better described as a random process than as a deterministic one. Norbert Wiener, the noted American mathematician, stated the case well

> ... In meteorology, the number of particles concerned is so enormous that an accurate record of their initial positions and velocities is utterly impossible; and if this record were actually made, and their future positions and velocities computed, we should have nothing but an impenetrable mass of figures which would need a radical reinterpretation before it could be of any service to us. The terms "cloud," "temperature," "turbulence," etc., are all terms referring not to one single physical situation but to a distribution of possible situations of which only one actual case is realized. If all the readings of all the meteorological stations on earth were simultaneously taken, they would not give a billionth part of the data necessary to characterize the actual state of the atmosphere from a Newtonian point of view. They would give only certain constants consistent with an infinity of different atmospheres, and at most, together with certain a priori assumptions, capable of giving us a probability distribution, a measure, over the set of possible atmospheres. Using the Newtonian laws, or any other system of causal laws whatever, all we can predict at any future time is a probability distribution of the constants of the system, and even this predictability fades out with the increase of time.
>
> -- Cybernetics, 1948

In this report, the weather is treated as a random process, and certain weather variables such as cloud cover, ceiling, and visibility are treated as random variables.

2.3 Random Variable

2.3.1 General. Let the set S represent the sample space of some experiment. The outcomes of the experiment constitute the sample points of S. Examples might be the number of heads in a series of coin tosses, the lifetime in hours of an electronic component, or the meteorological visibility in statute miles. These are all examples of random variables, i.e., functions whose value depends on the outcome of one or more chance events. The key point being made is that a random variable is not really a variable at all; it is a function.

Definition: A random variable X on a sample space S is a function or mapping from S into the set R of real numbers such that the preimage of every interval of R is an event of S.

The notions of image and preimage come from the underlying definition of a function. Let S and T be arbitrary sets with elements s and t

$$s \in S \tag{1}$$

$$t \in T \tag{2}$$

In addition, suppose that for each $s \in S$ there corresponds a unique element $t \in T$

s(1)	s(2)	s(3)	...	s(n)
t(1)	t(2)	t(3)	...	t(n)

The collection f of such mappings from S into T is called a function, written $f: S \rightarrow T$. In functional notation, when we write

$$f(s) = t \tag{3}$$

we are representing the element of T that the function f assigns to $s \in S$. That element is called the image of s under f or the value of the f at s. Let A be a subset of set S; then the image f(A) is defined by

$$f(A) = \{f(s): \; s \in A\} \tag{4}$$

where the elements of the set f(A) are defined by the expression in braces $\{\cdot\}$. Correspondingly, if B is a subset of T, then the preimage $f^{-1}(B)$ is defined by

$$f^{-1}(B) = \{s: \; f(s) \in B\} \tag{5}$$

In other words, f(A) consists of the images of points in A, and $f^{-1}(B)$ consists of those points whose images are in B. It is useful to note that the set f(S) of all the image points of S is called the image set or range of the function f.

A random variable X is the function X having the following properties

(1) The range of X is the set R of real numbers.[5]

[5] The range of a function is the set of values that the function takes on.

(2) The domain of X is contained in a set T certain of whose subsets correspond to events for which there is associated a probability function or distribution function.[6]

(3) For each real number x, the set of all $t \varepsilon T$ for which $X(t) \leq x$ is an event, i.e., has a probability, namely the probability that $X \leq x$.

The shorthand notation $Pr(X = a)$ can be used to represent the "probability that X maps into a" or $Pr(a \leq X \leq b)$ for the "probability that X maps into the closed interval [a,b]"

$$Pr(X = a) = Pr(\{s \varepsilon S: X(s) = a\}) \qquad (6)$$

$$Pr(a \leq X \leq b) = Pr(\{s \varepsilon S: a \leq X(s) \leq b\}) \qquad (7)$$

A random variable can be either discrete or continuous. If discrete, the function X(S) can take on only a finite number of values $x_1, x_2, x_3, \ldots, x_n$ with each of which there is an associated non-negative probability $Pr(X = x_i)$, the sum of all of which is unity. An example of a discrete random variable is the number of spots thrown with one die, which can take on the values 1, 2, 3, 4, 5, or 6, each of which has a probability of 1/6. If continuous, the function X(S) can take on any value in the set of real numbers R whereupon it becomes impossible to conceive of the probability of any particular value of X, and one must consider the probability of an interval of X. An example of a continuous random variable is the daily total rainfall.

In the discussion below, X will be a random variable, and x will be a particular value or possible value of X.

2.3.2 Probability Function of a Discrete Random Variable. Let X be a random variable on a sample space S with a finite or discrete image set, say

$$X(S) = \{x_1, x_2, \ldots, x_i, \ldots, x_{n-1}, x_n\} \qquad (8)$$

In other words, X is a real-valued, discrete random variable that takes on one of a finite number n of possible values x_i. X(S) can be mapped into probability space by defining the probability function (also called the distribution) $Pr(X = x_i) = P_X(x_i)$ as the probability that X will take on the particular value x_i.

[6] The domain of a function is the set of values that the independent variable takes on.

In the case of X representing the number of spots thrown from a single die, the probability function is as follows

x_i	1	2	3	4	5	6
$P_X(x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

The <u>cumulative distribution function</u> $F_X(x)$ for a discrete random variable is the sum of the probabilities of all x_i that are less than or equal to the threshold value x, i.e.,

$$F_X(x) = Pr(x_i \leq x) = \sum_{x_i \leq x} P_X(x_i) \tag{9}$$

Since discrete random variables are not used in the simulation models represented in this report, no further discussion of the discrete random variables and its probability functions is given here.

2.3.3 <u>Probability Density Function of a Continuous Random Variable</u>. Now let X(S) be a random variable on a sample space S with a continuous image set, i.e., the image set X(S) is a continuum of numbers such as the interval set $\{a \leq X \leq b\}$. Since the set $\{a \leq X \leq b\}$ is an event in X, it is possible to speak of the probability $Pr(a \leq X \leq b)$. This can be done through the mechanism of the integral calculus by introducing the concept of a <u>probability density function</u>, f_X.

Assume that a piecewise continuous function f_X exists such that the probability $Pr(a \leq X \leq b)$ is equal to the area under the graph of f_X between x = a and x = b, i.e.,

$$Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx \tag{10}$$

where X is the random variable and x is a dummy variable. The function f_X is called a <u>probability density function</u> of X and has "units" of "probability per unit X." The probability density function f satisfies the conditions that (1) f is non-negative and (2) the total area under its graph is unity, i.e.,

$$\int_R f_X(x)\, dx = 1 \tag{11}$$

The <u>cumulative distribution function</u> F_X of the continuous random variable X is defined as the probability that X will take on some value less than or equal to a threshold value x, i.e.,

$$F_X(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{12}$$

where t is a dummy variable. The cumulative distribution function satisfies the conditions that (1) F_X is monotonically increasing, i.e.,

$$F_X(a) \leq F_X(b) \qquad \text{for } a \leq b \tag{13}$$

and (2) the lower limit of F_X is zero, i.e.,

$$\lim_{x \to -\infty} F_X(x) = 0 \tag{14}$$

and (3) the upper limit of F_X is unity, i.e.,

$$\lim_{x \to -\infty} F_X(x) = 1 \tag{15}$$

It is apparent that the probability density function f_X of a continuous random variable X is the derivation of the <u>cumulative distribution function</u> F_X, i.e.,

$$f_X(x) = dF_X/dx \geq 0 \tag{16}$$

Note the relationship between probability and cumulative probability

$$\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) \tag{17}$$

$$= \int_{-\infty}^{b} f(t)\ dt - \int_{-\infty}^{a} f(t)\ dt \tag{18}$$

$$= F_X(b) - F_X(a) \tag{19}$$

The probability that a continuous random variable X takes on a single specified value d is zero, as can be seen from this analysis

$$\Pr(X=d) = \int_{d}^{d} f(t)\ dt = F_X(d) - F_X(d) = 0 \tag{20}$$

Since the probability that a continuous random variable takes on a particular value is zero,

$$\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a \leq X < b) = \Pr(a < X < b) \tag{21}$$

and

$$\Pr(X \leq x) = \Pr(X < x) \tag{22}$$

2.3.4 <u>Functions of a Random Variable</u>. Every function of a random variable is also a random variable. If X is a random variable, then

$$Z = g(X) \tag{23}$$

is a random variable as well.

2.3.5 Joint Probabilities of Continuous Random Variables. Let X and Y be continuous random variables whose joint probability density function is $f_{XY}(x,y)$. For these variables, the cumulative probability distribution is $F_{XY}(x,y)$. The two are related by

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y) \tag{24}$$

and the joint cumulative distribution function is

$$F_{XY}(x,y) = \Pr(X \leq x \text{ and } Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(s,t)\, ds\, dt \tag{25}$$

2.3.6 Marginal Distributions of Continuous Random Variables. A marginal probability distribution is the probability distribution of one variable regardless of the value of the other variable(s).

If X and Y are continuous random variables whose joint probability density function is $f_{XY}(x,y)$, and if one is interested only in the behavior of one of the variables, say X, then one can obtain $f_X(x)$, the marginal probability density function of X, by integrating the joint density function over all possible values of Y

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,s)\, ds \tag{26}$$

Marginal Probability Density of X	=	Y-integrated Joint Probability Density of X and Y

The cumulative marginal distribution is given by

$$F_X(x) = F_{XY}(x,\infty) = \Pr(X \leq x \text{ and } Y \leq \infty) \tag{27}$$

$$= \Pr(X \leq x) \tag{28}$$

$$= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(s,t)\, ds\, dt \tag{29}$$

$$= \int_{-\infty}^{x} f_X(s)\, ds \tag{30}$$

2.3.7 Conditional Probabilities of Continuous Random Variables. A conditional probability distribution is the distribution of one variable with restrictions or conditions placed on the second variable. For example, Pr(B|A) is the conditional probability of event B occurring given that event A has occurred or is occurring. Conditional probabilities can be expressed in terms of joint probabilities as follows

$$\Pr(B|A) = \Pr(A \cap B) \;/\; \Pr(A) \qquad (31)$$

where $\cap$ represents the intersection of events A and B, and therefore, $\Pr(A \cap B)$ is the joint probability of A <u>and</u> B.

Consider two continuous random variables X and Y whose joint probability density is $f_{XY}(x,y)$. It might be of interest to know the conditional distribution of X given that Y is in some region R, e.g., R: $y_1 \leq Y \leq y_2$.

Following Equation (31) above, but using probability densities instead of probabilities, one can write

$$\text{Probability Density}(x|Y \text{ in } R) =$$

$$\frac{\text{Y-integrated Joint Probability Density } (x \cap Y \text{ in } R)}{\text{Marginal Probability Density}(Y \text{ in } R)} \qquad (32)$$

Notation for the probability density of x given Y in R is

$$f_{X|Y}(x|Y \text{ in } R)$$

The marginal probability density of Y in R is obtained by integrating the joint density $f_{XY}(x,y)$ over all X, i.e.,

$$\int_{-\infty}^{\infty} \int_{R} f_{XY}(s,t)\, ds\, dt = \int_{R} f_{Y}(t)\, dt \qquad (33)$$

The Y-integrated joint probability density of X is

$$\int_{R} f_{XY}(x,t)\, dt$$

The equation for the conditional probability density of X given Y in R is thus

$$f_{X|Y}(x|Y \text{ in } R) = \frac{\int_R f_{XY}(x,t)\, dt}{\int_{-\infty}^{-\infty} \int_{R} f_{XY}(s,t)\, ds\, dt} \qquad (34)$$

$$= \frac{\int_R f_{XY}(x,t)\, dt}{\int_R f_{Y}(t)\, dt} \qquad (35)$$

Sometimes it is desired to find the conditional probability density of X given that Y is equal to some particular value y_0. In other words, the region R reduces to the point y_0, and an argument in the limit leads to the result,

$$f_{X|Y}(x|Y=y_0) = f_{XY}(x,y_0) \;/\; f_Y(y_0) \tag{36}$$

The conditional cumulative probability distribution function of X given that Y has taken on the particular value y_0 is therefore

$$F_{X|Y}(x,y) = Pr(X \leq x|Y = y_0) = \int_{-\infty}^{x} f_{X|Y}(s|Y = y_0)\, ds \tag{37}$$

$$= \int_{-\infty}^{x} f_{XY}(s,y_0) \;/\; f_Y(y_0)\, ds \tag{38}$$

2.3.8 Independence. In general, the conditional probability density function of X given Y is a function of the value y taken on by Y. If the random variables X and Y are independent, then the probability of X does not depend on Y, and the conditional probability density of X given Y reduces to the marginal density of X alone, i.e.,

$$f_{X|Y}(x|y) = f_X(x) \tag{39}$$

Furthermore, in the case of independent X and Y, the joint probability density of X and Y is equal to the product of the marginal densities

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) \tag{40}$$

2.3.9 Expectation. Let X be a continuous random variable whose probability density function is $f_X(x)$. Let g be a real valued function of X. Then the expectation or expected value of the function g is defined as

$$E[g(x)] = \int_{R} g(x)\, f_X(x)\, dx \tag{41}$$

$$= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \tag{42}$$

where $E[\cdot]$ is called the expectation operator, and R is the set of real numbers.

In the case where

$$g(x) = X \tag{43}$$

the expectation of X is defined

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \tag{44}$$

One important property of the expectation operator is that the expectation of a linear function of X is a linear function of the expectation of X. This can be seen by considering

$$g(X) = a + bX \tag{45}$$

where a and b are constants. Taking the expectation of g(X) yields

$$E[g(X)] = \int_{-\infty}^{\infty} [a + bX]\ f_X(x)\ dx \tag{46}$$

$$= a \int_{-\infty}^{\infty} f_X(x)\ dx + b \int_{-\infty}^{\infty} X\ f_X(x)\ dx \tag{47}$$

$$= a + b\ E[X] \tag{48}$$

because $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$. Thus,

$$E[a + bX] = a + bE[X] \tag{49}$$

leading to the result

$$E[a] = a \tag{50}$$

or the expectation of a constant is a constant.

Let X and Y be continuous random variables whose joint probability density function is $f_{XY}(x,y)$. Let g(X,Y) be a function of the two random variables. Then the expectation of the function g is

$$E[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\ f_{XY}(x,y)\ dx\ dy \tag{51}$$

2.3.10 Correlation. *Correlation is a measure of association, not of causation.* Loosely, we can say that correlation, and in particular the various correlation "coefficients," are measures of relatedness among variables. But statistical relatedness does not necessarily imply physical causation.

In meteorology, correlations can sometimes arise from causal relationships and sometimes from other sources such as covariation, biased data, and artificial correlation introduced by using derived or functionally related variables in the analysis at hand. As an example of covariation in meteorology, one can cite the high positive correlation between the low-level moisture in Georgia and the occurrence of afternoon thunderstorms. There is also a high positive correlation between low-level moisture and morning fog. These two correlations, for which physical causation could be strongly argued, give rise to another correlation: between the occurrence of fog and the occurrence of thunderstorms. But does that correlation, however strong it may be, imply that morning fog causes afternoon thunderstorms, or that afternoon thunderstorms cause the previous morning's fog? Probably not. The relation between fog and thunderstorms is through a third variable (often called the "covariate") common to both. In this case, the covariate is low-level moisture, and the "true" relationship between morning fog and afternoon thunderstorms can only be estimated by isolating the role played by the covariate, low-level moisture.

Correlation, then, is a way of expressing the association between variables, an association that need not be causal in nature.

So far, our ideas about correlation have been expressed qualitatively. In statistics, however, the "correlation coefficient" or sometimes just the "correlation" is a quantitative concept, capable of being expressed in numbers that describe the degree of relatedness or association between variables. In earlier sections of this chapter, there is scale for measuring probability; similarly, a scale for measuring correlation is desirable. When there is no relationship between variables, this statistical measure ought to approach zero. The measure of correlation should approach unity when the relationship between variables is very high. While there is no such thing as negative probability, it is easy to have a negative correlation. For example, there may be two variables, A and B, that increase together (positive correlation), or two variables, C and D, one of which increases as the other decreases (negative correlation).

A general definition of correlation can be set down: two measurable characteristics, A and B, are said to be correlated when, with different values x of A, the same value y of B is not equally likely to be associated. In other words, certain values of B are more likely to occur with the value x than others. If they were not, correlation would be absent. Correlation would be perfect if for every value of A the same value of B occurred.

The correlation coefficient measures the relative importance of the relationship between two variables in a nondimensional sense, i.e., it does not depend upon any arbitrary choice of units by which the original variables were measured. The concept of a theoretical population correlation coefficient can be developed along the following lines.

Let X and Y be two random variables on a sample space S such that

$$X(S) = \{x_1, x_2, \ldots, x_n\} \tag{52}$$

$$Y(S) = \{y_1, y_2, \ldots, y_n\} \tag{53}$$

with joint probability density function $f_{XY}(x,y)$. Then the covariance of X and Y, denoted by Cov(X,Y), is defined by

$$\mathrm{Cov}(X,Y) \equiv E[(X - \mu_X)(Y - \mu_Y)] \tag{54}$$

$$= E(XY) - E(X)E(Y) \tag{55}$$

$$= E(XY) - \mu_X \mu_Y \tag{56}$$

where μ_X is the mean of X and μ_Y is the mean of Y. Assuming a linear relationship between the random variables X and Y results in the following expression for the theoretical linear correlation coefficient between X and Y

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \, \sigma_Y} \qquad \text{(Linear)} \qquad (57)$$

where σ_X is the standard deviation of X and σ_Y the standard deviation of Y.

Equation (57) is an expression for the theoretical linear correlation coefficient. For problems requiring sampling of actual data, it is not the theoretical correlation coefficient ρ but rather the sample correlation coefficient r that is of interest.

An expression for the sample correlation coefficient r_{XY} can be developed by considering a set of (X,Y) data pairs, where Y is considered a function of X

$$\hat{Y} = f(X) \qquad \text{(General)} \qquad (58)$$

Here Y are actual values of the dependent variable, and $\hat{Y}$ are the Y-values predicted by the function f. If f is a linear function, then Equation (58) particularizes to

$$\hat{Y} = a_0 + a_1X \qquad \text{(Linear)} \qquad (59)$$

The linear function f in one independent variable may not perfectly describe the behavior of the Y-data. There may, for example, be independent variables other than X that are important in predicting Y, or there may be a nonlinear dependence involved. The scatter of actual Y-values about the prediction $\hat{Y}$ given by Equation (59) can be described in terms of the standard error of the estimate of Y on X, given by

$$s_{YX} = \sqrt{\Sigma[(Y - \hat{Y})^2] \;/\; N} \qquad \text{(General)} \qquad (60)$$

which applies to both linear and nonlinear associations of the form shown in Equation (58). If the linear association of Equation (59) is used, then Equation (60) becomes

$$s_{XY}{}^2 = \frac{\Sigma Y^2 - a_0\Sigma Y - a_1\Sigma XY}{N} \qquad \text{(Linear)} \qquad (61)$$

a measure of the standard error of the linear estimate of Y on X.

The total variation of Y is defined as $\Sigma(Y - \bar{Y})^2$, the sum of the squares of the deviations of Y from its mean $\bar{Y}$. That total variation of Y can be partitioned into an unexplained variance $\Sigma(Y - \hat{Y})^2$ and an explained variance $\Sigma(\hat{Y} - \bar{Y})^2$

$$\underset{\text{Total Variation}}{\Sigma(Y - \bar{Y})^2} = \underset{\text{Unexplained Variation}}{\Sigma(Y - \hat{Y})^2} + \underset{\text{Explained Variation}}{\Sigma(\hat{Y} - \bar{Y})^2} \qquad \text{(General)} \qquad (62)$$

where $\hat{Y}$ is an estimated value of Y based on the value of X and the functional relationship expressed by Equations (58) and (59). The first term on the right of Equation (62) is called the "unexplained" variation because the deviations behave in an apparently random or unpredictable manner. The second term on the right of Equation (62) is called the "explained" variation because the deviations involved have a definite pattern.

The sample <u>correlation coefficient</u> r_{XY} between the variables X and Y is given by

$$r_{XY} = \pm \sqrt{\frac{\text{Explained Variation}}{\text{Total Variation}}} = \pm \sqrt{\frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}} \qquad \text{(General)} \qquad (63)$$

which varies between -1 (perfect negative correlation) and +1 (perfect positive correlation) and is nondimensional and independent of the origin. The ± sign is used to introduce the sign of the correlation. Equation (63) is a perfectly general expression for the correlation coefficient and can be used for linear or nonlinear correlation. Using Equation (60) in (62), and making use of the fact that the standard deviation of Y is

$$s_Y = \sqrt{[\Sigma(Y - \bar{Y})^2] \; / \; N} \qquad (64)$$

permits Equation (63) to be written as

$$r_{XY} = \sqrt{(s_Y{}^2 - s_{YX}{}^2) \; / \; s_Y{}^2} \qquad \text{(General)} \qquad (65)$$

Like Equation (63), Equation (65) is perfectly general and can be used for nonlinear as well as linear correlation. If $\hat{Y}$ is computed from a nonlinear function (Equation 58), and the ± signs are omitted, then Equations (63) and (65) describe nonlinear correlation.

$$r_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2} \; \sqrt{\Sigma(Y-\bar{Y})^2}} \qquad \text{(Linear)} \qquad (66)$$

where the covariance of X and Y is

$$s_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N} \tag{67}$$

and the standard deviations are given by Equation (64) for s_Y and by its analogue for s_X.

If Equations (67), (64), and the analogue mentioned immediately above are used in Equation (66), the result is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \qquad \text{(Linear)} \tag{68}$$

which parallels Equation (57) for the linear population correlation ρ_{XY}.

The interpretation attached to the sample correlation coefficient r_{XY} depends on the functional form introduced in Equation (58) for the association between the two variables X and Y. If a linear association is assumed, then r_{XY} is calculated from Equation (66) and measures the extent to which a linear dependence of Y on X explains the variation of Y data, and $r_{XY}{}^2$ becomes the fraction of the total variation of Y explained by a linear dependence on X. If a nonlinear association is used for Equation (58), then $r_{XY}{}^2$ -- which can then no longer be calculated from Equation (66) -- is the fraction of the total variation of Y explained by a particular nonlinear association with X. Just because there is no linear correlation between the variables X and Y does not mean there is no correlation at all. There may in fact be a high nonlinear correlation between the variables.

The stochastic process models developed in this report and applied to the task of environmental simulation modeling employ linear correlation methods exclusively. Therefore, throughout the remainder of this report, all references to correlation will refer to linear correlation.

USAFETAC uses two methods most frequently when calculating linear correlation coefficients: (1) the Pearson product moment (PPM) formula (Equation 66 above), and (2) the tetrachoric method.

The Pearson product-moment (PPM) formula for calculating linear correlation is basically Equation (66). That form of the equation is computationally inefficient because it requires the means to be known in advance (from an earlier pass through the data). Algebraic manipulation of Equation (66) produces a computationally efficient PPM formula

$$r_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X2-(\Sigma X)^2]}\,\sqrt{[N\Sigma Y^2-(\Sigma Y)^2]}} \tag{69}$$

The correlation coefficient r_{XY} is symmetrical in X and Y. This symmetry indicates that the coefficient of correlation does not distinguish between the dependent and the independent variable. It permits conclusions about the existence of a linear relationship between two variables but not about which "depends" on the other. Although it is relatively easy to implement on a computer, Pearsons' method has the inherent disadvantage of requiring the raw data to be available for the computations. The tetrachoric method offers the advantage of being able to use data that has already been categorized.

Any two variables can be reduced to a two-by-two table

		X	
		Above	Below
		X_t	
Y	ABV	A	B
Y_t	BLW	C	D

where A, B, C, and D are the number of cases above or below the critical, or threshold values, (X_t and Y_t) of the respective variables. An approximation to the tetrachoric correlation coefficient (r_t) can be obtained by Equation (70)

$$r_t = \sin \left[\frac{\pi}{2} \frac{\sqrt{AD} - \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}} \right] \qquad (70)$$

This equation is accurate where (A + B)/N and (A + C)/N are close to 0.5 (where N is the total number of cases), but may contain sizable error for values near one or zero. Since there is no simple exact formula for calculating r_t, an algorithm based on the false position method (Acton, 1970) is used by USAFETAC. The coefficient is evaluated at two initial guess values, and linear interpolation is used to find a better estimate. The quantity r_t behaves in a manner similar to an ordinary linear correlation coefficient, but the exact numerical value is not completely comparable. The value of r_t varies from -1 to +1, giving zero for no relation, but the sign depends in a rather arbitrary manner on the arrangement of the contingency table.

2.4 Stochastic Processes

Parzen (1962) points out that the term stochastic is of Greek origin, that in 17th century English the word meant "to conjecture" or "to aim at a mark," and that today the word has come to mean "pertaining to chance." In modern practice, the words stochastic, random, and chance are used as synonyms.

A stochastic process or random process is a succession of values taken on by a random variable X(t) as a function of the parameter $t \varepsilon T$. The set T is called the index set of the process. Random processes are controlled by probabilistic laws. In many applications, the index parameter t of the stochastic process represents time but also can be used as some sort of event sequence number.

From the point of view of mathematical statistics, a stochastic process is best defined as the collection

$$\{X(t),\ t \varepsilon T\}$$

of random variables X(t) all defined on the same sample (probability) space. No restriction is placed on the nature of the index set T, but two important cases arise from the nature of T

- Discrete Parameter Process: T is a countable set $T = \{0, \pm 1, \pm 2, \ldots\}$ or $T = \{0, 1, 2, \ldots\}$.

- Continuous Parameter Process: T is an uncountable subset of the set R of real numbers, so $T = \{t: -\infty < t < \infty\}$ or $T = \{t: \geq 0\}$.

A time series is a finite realization of a stochastic process where the index parameter t represents time. A time series can be produced either in the form of output from a model or in the form of experimental data. A time series, in other words, is a sequence of values of a random variable collected over discrete or continuous time.

In a stochastic process model, a random variable q_t can be formed as the sum

$$q_t = d_t + \varepsilon_t \tag{71}$$

of a deterministic part d_t and a random or stochastic part ε_t. Typically, the deterministic part contains the contribution of preceeding values q_{t-1}, q_{t-2}, etc., in the series but may also have terms such as $\bar{q}$ representing the mean value or $\tilde{q}$ representing a secular or long-term trend in values. The random part ε_t of the solution introduces noise or uncertainty into the process being modeled; otherwise, the process would not be random at all. As shown in Equation (71), there is no restriction on the form of ε_t, but in practice it tends to be either (1) a number drawn at random from a population distributed uniformly over the interval [0,1] with mean of 1/2 and variance of 1/12, or (2) a number drawn at random from a population distributed normally over the interval $(-\infty,\infty)$ with mean of zero and variance of 1, i.e., N(0,1). That is to say, ε_t is either a uniform random number or a normal random number.

If the stochastic process shown in Equation (71) above is further assumed to be covariance-stationary, then neither the mean(s) nor the variance(s) of the quantit(ies) being simulated are dependent on the index parameter t (i.e., they do not change with time if t represents time), and the covariance between two successive values

$$q_t \quad \text{and} \quad q_{t+\Delta t}$$

i.e., $\mathrm{Cov}(q_t, q_{t+\Delta t})$ becomes a function only of the separation Δt between the two, and does not depend on the absolute values of the index parameter t. Also, the correlation ρ between successive values of q becomes dependent only on the separation Δt, i.e.,

$$\rho_{t,t+\Delta t} = \rho_{\Delta t} = \frac{\mathrm{Cov}(q_t, q_{t+\Delta t})}{\sqrt{\sigma_t^2 \; \sigma_{t+\Delta t}^2}} = \frac{\mathrm{Cov}(q_0, q_{\Delta t})}{\sigma^2} \tag{72}$$

(because $\sigma_t^2 = \sigma^2 = \sigma_{t+\Delta t}^2$).

Applying the covariance-stationary assumption to the process of Equation (71) leads to the linear autoregressive (AR) relation,

$$q_t = \beta_0 + \beta_1 q_{t-1} + \beta_2 q_{t-2} + \cdots + \beta_m q_{t-m} + \varepsilon_t \tag{73}$$

where the β_i are the autoregression coefficients, and the ε_t is an independent error term. In this formulation, the deterministic part of the solution depends on the lag-one value q_{t-1}, the lag-two value q_{t-2}, etc., and the random part of the solution is now an independent error term with mean of zero.

The AR process (Equation 73) can be further restricted by applying the first-order Markovian assumption that the value q_t of the process at t depends on the previous value q_{t-1} alone, not on how the process reached q_{t-1}. Then the model becomes

$$q_t = \beta_0 + \beta_1 q_{t-1} + \varepsilon_t \tag{74}$$

an autoregressive (AR), first-order Markov model. For such models, the serial correlation ρ (the correlation in the t-dimension) follows an exponential decay law (see Appendix C)

$$\rho_{\Delta t} = \rho_1^{\Delta t} \tag{75}$$

where ρ_1 is the serial correlation for lag $\Delta t = 1$. Equation (75) shows that for a Markov model, realizations spaced Δt units apart will have correlation $\rho_1^{\Delta t}$.

In order to estimate the parameters β_0 and β_1 and to specify the form of the error term ε_t, it is helpful to assume that q_t and q_{t-1} are derived jointly from a bivariate normal population with means

$$\mu_t = \mu_{t-1} = \mu \tag{76}$$

and variances

$$\sigma_t{}^2 = \sigma_{t-1}{}^2 = \sigma^2 \tag{77}$$

This causes the regression function of q_t on q_{t-1} to be linear and homoscedastic (of constant variance). The conditional expectation of q_t given q_{t-1} is

$$E(q_t | q_{t-1}) = \mu + \rho(q_{t-1} - \mu) \tag{78}$$

where ρ is the correlation between q_t and q_{t-1} and where

$$\mathrm{Var}(q_t | q_{t-1}) = \sigma^2(1 - \rho^2) \tag{79}$$

which is independent of q_{t-1}.

As shown in almost any elementary statistics text, the standard normal variable (equivalent normal deviate) z_w corresponding to the normally distributed raw variable w with mean (expected value) μ_w and standard deviation σ_w is

$$z_w = \frac{w - \mu_w}{\sigma_w} \tag{80}$$

Hence, the value of the raw variable w can be calculated from

$$w = \mu_w + \sigma_w z_w \tag{81}$$

Using $q_t | q_{t-1}$ for w in Equation (81), and substituting from Equation (78) for μ_w and from Equation (79) for σ_w yields

$$q_t = \mu + \rho(q_{t-1} - \mu) + \sigma\sqrt{1-\rho^2}\ z_q \tag{82}$$

where z_q is a random normal number.

Comparing Equation (82) with (74) shows that

$$\beta_0 = \mu(1 - \rho) \tag{83}$$

$$\beta_1 = \rho \tag{84}$$

$$\varepsilon_t = \sigma\sqrt{1 - \rho^2}\ z_q \qquad (85)$$

Equation (82) can be rearranged and transformed into mean-deviation form using

$$v_{(\cdot)} = q_{(\cdot)} - \mu_{(\cdot)} \qquad (86)$$

with the result,

$$v_t = \rho v_{t-1} + \sigma\sqrt{1 - \rho^2}\ \eta \qquad (87)$$

where η, like z_q, is simply a random normal number.

Equations (71), (73), and (74), (82), and (87) present a spectrum of increasingly more specific and more restrictive stochastic process models. Equation (71) is a very general form that can describe almost any stochastic process model. Equation (73) represents a covariance-stationary, linear autoregressive (AR) process. Finally, Equations (74), (82), and (87) further require the first-order Markov assumption be made and normally distributed random variables used. Equations (82) and (87) require that the variables q_t and q_{t-1} or v_t and v_{t-1} be jointly distributed according to the bivariate normal probability distribution. Equation (87) is the Ornstein-Uhlenbeck stochastic process model that forms the basis for much of the present work in environmental simulation modeling.

2.5 Markov Processes

2.5.1 Introduction. In classical meteorology, just as in classical physics, deterministic laws are set forth in the form of initial value problems, in which, given the state of the atmosphere at some initial time t_0, it is possible to deduce its state at a later time t_1. In this formulation, remembrance of the state of the atmosphere at any time prior to t_0 is irrelevant to the question of deducing the state at t_1. Numerical weather prediction models, which express the *evolution of meteorological mass and motion fields in terms of classical* deterministic physical principles, are "memoryless" initial value problems of this sort.

Just as in physics, where phenomena such as radioactive decay and Brownian motion have had to be described probabilistically, so also in meteorology have been described such smaller scale phenomena as turbulent motions and hourly changes in ceiling, visibility, sky cover, and wind speed at a point. For physical quantities whose behavior is best treated in terms of probabilistic laws rather than deterministic ones, there exists a "memoryless" formulation analogous to the initial value problem of classical mathematical physics. This is the so called first-order Markov process, in which the probability that a physical system will be in state x_1 at time t_1 may be deduced strictly from knowledge of the system's state x_0 at time t_0 and does not depend on the history of the system before t_0, i.e., the system's state at t_1 depends only on the state at t_0 and not on the path by which the state at t_0 was reached.

Markov processes are classified according to (1) the nature of the state space {X} of the process, and, (2) the nature of the index set T or parameter t of the process.

2.5.2 Discrete-state vs. Continuous-state Markov Processes. In both the cases presented above, the state space {X} of the process was taken as discrete valued, i.e., $\{X\} = \{x_k, k = 1, 2, \ldots K\}$. Markov processes whose state space is discrete valued are called Markov chains. It is also possible to describe stochastic processes in general and Markov processes in particular whose state space {X} is continuous. In such a continuous case, a real number x is said to be a possible value or state of the stochastic process {X(t)} if there exists a time t such that the probability,

$$\Pr\{x-h < X(t) < x+h\}$$

is positive for every $h > 0$.

2.5.3 Discrete-parameter vs. Continuous-parameter Markov Processes. Mathematically, a Markov process can be defined as either a discrete parameter stochastic process or a continuous parameter stochastic process, depending on whether its index parameter t is discrete or continuous. A discrete parameter stochastic process can be expressed as the set of random variables, $\{X(t), t = t_0, t_1, t_2, \ldots, t_n\}$. A continuous parameter stochastic process can be expressed as the set $\{X(t), t \geq 0\}$. In a first-order, discrete parameter Markov process, the conditional probability of $X(t_n)$ depends only on $X(t_{n-1})$, the most recent known value, i.e.,

$$\begin{aligned}\Pr\{X(t_n) \leq x_n \mid X(t_1) = x_1, X(t_2) = x_2, \ldots X(t_{n-1}) = x_{n-1}\} \\ = \Pr\{X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}\} \qquad (88)\end{aligned}$$

The discrete parameter Markov process $\{X_n\} = \{X(t_n)\}$ for parameter t given by $n \geq m > 0$ and states $x_j = j$ and $x_k = k$ is described by the probability mass function,

$$p_j(n) = \Pr\{X_n = j\} \qquad (89)$$

and the conditional probability mass function,

$$p_{j,k}(m,n) = \Pr\{X_n = k \mid X_m = j\} \qquad (90)$$

The function $p_{j,k}(m,n)$ is called the transition probability function of the Markov process.

In the case where $\{X(t),\ t \geq 0\}$ is a continuous parameter Markov process, then the process is defined for all index values $t \geq s \geq 0$ and states j and k by the probability mass function,

$$p_k(t) = \Pr\{X(t) = k\} \tag{91}$$

and the conditional probability mass function,

$$p_{j,k}(s,t) = \Pr\{X(t) = k \mid X(s) = j\} \tag{92}$$

The function $p_{j,k}(s,t)$ is again called the transition probability function of the Markov process.

2.5.4 <u>Order of the Markov Process</u>. Strictly speaking, the "memoryless" Markov process discussed above, in which the conditional probability of the state $X(t_n)$ depends only on the immediately preceding state $X(t_{n-1})$ is called a first-order Markov process. It is possible to define a second-order Markov process, in which the conditional probability of $X(t_n)$ would depend not only on $X(t_{n-1})$ but also on $X(t_{n-2})$, and in which the transition probability matrices are three-dimensional. Even higher order Markov processes can be easily conceived, if not so easily understood and applied.

When applying Markov models to data, one of the tasks at hand is to estimate the <u>order</u> of the Markov model that best fits the data.

2.5.5 <u>Relationship to Autoregressive-Moving Average (ARMA) Models</u>. The set of Markov models is a subset of the very flexible family of autoregressive-moving average (ARMA) models, sometimes called Box-Jenkins models (Box and Jenkins, 1976). In general, an ARMA model takes the form,

$$z_{t+1} = \underbrace{\sum_{i=1}^{p} \phi_i z_{t-i+1}}_{\substack{\text{Autoregres-}\\ \text{sive Terms}\\ \text{(AR)}}} + \underbrace{v_t}_{\substack{\text{Current}\\ v_t}} - \underbrace{\sum_{j=1}^{p} \theta_j v_{t-j}}_{\substack{\text{Moving Aver-}\\ \text{age Terms}\\ \text{(MA)}}} \tag{93}$$

where Z is a random variable with mean of zero and where the V values represent independent, identically distributed random variables having the normal, or N(0,1), distribution. Following convention, z is a particular value of the random variable Z, and v is a particular value of V.

The model described in Equation (93) is an ARMA model of order (p,q), having p autoregressive (AR) and q moving average (MA) terms. If only the moving average terms are retained, the model becomes ARMA(0,q), i.e., an MA(q) model. If only the autoregressive terms are retained, the model becomes ARMA(p,0), or

AR(p). AR(p) models are Markov models of order p. Consider an AR(1) model, which must be first-order Markov

$$z_{t-1} = \phi_1 z_t + v_t \tag{94}$$

Comparing Equation (94) with the Ornstein-Uhlenbeck first-order Markov model expressed in Equation (87) shows that such a model is actually an AR(1) model in which $\phi_1 = \rho$ and in which $v_t = \sigma\sqrt{1-\rho^2}\,\eta$, i.e., a special case of the ARMA(p,q) model.

Estimating the order of an ARMA model consists in finding the values of p and q for which the model best fits the data, which are usually in the form of a time series. This is done by calculating the autocorrelations r_k for lag k and the partial autocorrelations ϕ_{kk} (see Appendix D) of the observed time series, and using these values as guidance for how many AR and MA terms <u>might</u> be needed in the eventual model. Then, based on that guidance, one actually fits recommended ARMA(p,q) models to the data, obtaining maximum likelihood estimators of the ARMA parameters ϕ_i and θ_j. Finally, one uses statistical tests to determine whether the ϕ_i and θ_j are significantly different from zero. In general, one's objective is to identify the simplest ARMA model that adequately describes the data.

Elements of the process of fitting ARMA (and especially AR(1) or first-order Markov) models to data are shown in the following paragraph.

2.5.6 <u>Fitting a Markov Model to Data: An Example Using the Wind Speed</u>. Consider a physical system that obeys some probabilistic law, such as wind observations on a mountaintop. The variable X can be used to represent the outcome of periodic observations of the system, so x_1, x_2, ..., x_i represent the first, second, and ith observations of the system. A possible sample after ten observations of wind velocity in meters per second, taken an hour apart, might be (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}) = (2, 1, 1, 3, 2, 6, 8, 7, 9, 9), as illustrated in Figure 1.

The order of the ARMA model that best fits the observed time series in Figure 1 is not readily apparent from casual inspection of the information in the figure. Many subjective arguments could plausibly be advanced, having to do with the apparent variability of the sample, the dependence apparent in physical processes, etc. These arguments could just as well lead to one conclusion or another regarding the order of the ARMA model of best fit.

Rather than to use informal visual inspection or subjective arguments to estimate the order of the ARMA model, it is better to use the statistical character of the time series data themselves to provide an estimate of the model that best fits those data. A problem arises immediately in that Figure 1 has only 10 data points, far too little information for any meaningful statistical study.

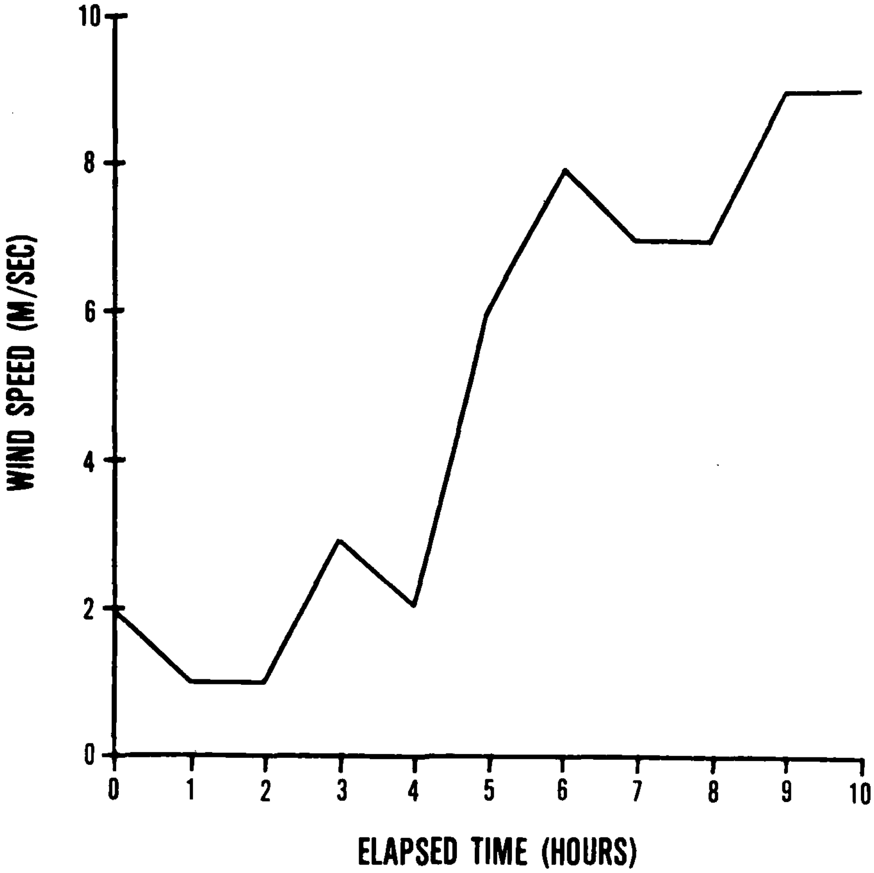


Figure 1. Results of Observations of Wind Speed on a Hypothetical Mountaintop.

For purposes of expediting the present discussion, it is convenient for now to ignore this problem of sampling error, assuming the data are sufficient in number to proceed with the statistical tests.

Autocorrelations r_k actually calculated from the Figure 1 data are 0.95, 0.93, 0.91 and 0.84 for k = 1, 2, 3, and 4 hours' lag, respectively. This is roughly an exponential decay of autocorrelation r_k as a function of lag k. Such behavior of r_k argues strongly for an ARMA(1,0) model but also suggests an ARMA(1,1) model. To distinguish between these, it is necessary to look at the partial autocorrelations ϕ_{kk} as a function of lag k. Values of ϕ_{kk} for k = 1, 2, and 3 are 0.95, 0.24, and 0.14, respectively. Confidence limits about the partial autocorrelations show that only the first, ϕ_{11}, is significantly different from zero. This condition argues strongly for an ARMA(1,0) or AR(1) first-order Markov model.

Rather than to fit an explicit Box-Jenkins ARMA model to the wind data of Figure 1, it is possible to transform the parameter t (representing time in this example) and the variable w (representing wind speed in this example) to discrete variables (by rounding to the nearest integral value, for example). If this is done, the wind speed model can be cast into the form of a Markov chain.

The hourly wind speed is represented by a discrete random variable W(t) in hour t, which takes on values $w_i(t)$ with unconditional probabilities p_i where,

$$\sum_{i=1}^{n} p_i = 1 \tag{95}$$

The value of W_{t+1} is not totally independent of W_t, especially at short time intervals. Such dependence can be modeled by a Markov chain. This requires specification of the transition probabilities,

$$p_{ij} = \Pr(W_{t+1}=w_j | W_t=w_i) \tag{96}$$

A transition probability is the conditional probability that the next wind speed state is w_j, given that the current wind speed is w_i. The transition probabilities satisfy

$$\sum_{j=1}^{n} p_{ij} = 1 \qquad \text{(for all i)} \tag{97}$$

The one-step transition probabilities can be arranged in matrix form

$$\underline{P} = \begin{vmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ p_{12} & p_{22} & \cdots & p_{2j} \\ \cdots & \cdots & \cdots & \cdots \\ p_{i1} & p_{i2} & \cdots & p_{ij} \end{vmatrix} \tag{98}$$

where $\underline{P}$ is the transition matrix whose elements are p_{ij}. For a Markov chain, the transition matrix contains all the information necessary to describe the behavior of the system. Let $p_i(t)$ be the probability that the system resides in state i at time t. Then the probability that $W_{t+1} = w_j$, is the sum of the probabilities $p_i(t)$ that $W_t = w_i$, times the probability p_{ij} that $W_{t+1} = w_j$, given that $W_t = w_i$.

$$p_j(t+1) = \sum_{i=1}^{n} p_i(t)\, p_{ij} \tag{99}$$

Letting $\underline{p}(t)$ be the row vector of state resident probabilities ($p_1(t)$, $p_2(t)$,..., $p_n(t)$), the relationship may be written,

$$\underline{p}(t+1) = \underline{p}(t)\, \underline{P} \tag{100}$$

It is then possible to compute the probabilities of each wind velocity for successive time intervals t+2, t+3, etc.

In summary, a Markov process is a probabilistic model for a continuous physical system from which a sample (x_1, x_2, ..., x_n) is available. The Markov process is characterized by the fact that the state of the system at time t_1 depends only on the state observed at time t_0. A Markov chain is a discrete approximation to a continuous process and completely describes the system when the state at time t_0, the initial probability vector $\underline{p}(t)$, and the one-step transition matrix $\underline{P}$ are given. In practice, USAFETAC uses a continuous form of the first-order Markov process, namely the Ornstein-Uhlenbeck model, which will be described in the next chapter.

Chapter 3

BASIC SINGLE-STATION MODELS

3.1 Single-variable, Single-station Model (V1S1)

USAFETAC's basic environmental simulation model is an Ornstein-Uhlenbeck stochastic process. This single-variable model is an autoregressive (AR), first-order Markov process in which each value of a random variable X_t is taken to be a particular value of a stationary stochastic process. It is a common and usually justifiable assumption to treat weather variables as a first-order Markov process (see Sections 2.4 and 2.5 of this technical note). The Ornstein-Uhlenbeck process is well based in the statistical literature and can be applied with substantial justification to variables whose time series have a random component and approximately adhere to the first-order Markov restriction.

The generation of a time series of a single meteorological variable would be quite simple if each value in the time sequence were independent of all others in the sequence. In general, this is not the case. Whether successive meteorological observations are independent depends on the time separation between them. The common separation between surface meteorological observations is 1, 3, or 6 hours. At these separations, successive observations of most meteorological variables are not serially independent. A goal of a simulation model should therefore be to reproduce this serial dependence between successive values of the particular meteorological variable being simulated, as well as to reproduce its probability distribution.

Assume that the variable to be simulated is normally distributed. If the variate is not normally distributed, then it can be transformed to the normal distribution by expressing the values of the raw variable in terms of its equivalent normal deviate (END). (Transformation of variables to the normal distribution is covered in detail in Boehm (1976) and is summarized in Section 3.1.1). The joint normal density function of two weather variables X_t and X_{t+1} at times t and t+1 with mean μ, variance σ^2, and serial correlation ρ between successive values is

$$f_{X_t X_{t+1}}(x_t, x_{t+1}) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{1/2}} \exp\left[\frac{(x_t-\mu)^2-2\rho(x_t-\mu)(x_{t+1}-\mu)+(x_{t+1}-\mu)^2}{2\sigma^2(1-\rho^2)}\right] \quad (101)$$

So the joint normal probability of two random variables with the same mean and variance depends only on μ, σ^2, and their correlation ρ. The generation of a time series of observations then requires the conditional distribution of the weather variable at one time given the value of the variable in previous hours If the weather process approximates a first-order Markov process, then the dependence of the distribution at time t+1 on the distribution of the variable in previous hours is summarized by the value of the variable at time t. If successive observations of this arbitary weather variable have a multivariate normal distribution, then the conditional distribution of X_{t+1} is normal with mean and variance equal to

$$E[X_{t+1}|X_t=x_t] = \mu + \rho(x_t-\mu) \tag{102}$$

$$Var[X_{t+1}|X_t=x_t] = \sigma^2(1-\rho^2) \tag{103}$$

where x_t is the value of X_t at hour t. This relationship is illustrated in Figure 2. From Equation (103) it can be seen that the larger the absolute value of the serial correlation ρ between the values of the variable, the smaller the conditional variance of X_{t+1}, which does not depend at all on the value of x_t.

A time series of synthetic, normally distributed variables with mean μ, variance σ^2, and a serial correlation ρ is produced by the equation,

$$X_{t+1} = \mu + \rho(X_t - \mu) + \sigma\sqrt{1-\rho^2}\,\eta_t \tag{104}$$

where η_t is a standard normal random number, i.e., a number drawn at random from a population with a mean of zero and a variance of unity, abbreviated as N(0,1). Each η_t is totally independent of past values of η as well as past values of X. If the variable being simulated is expressed as an END (which itself is distributed N(0,1)), then Equation (104) simplifies to

$$X_{t+1} = \underset{(a)}{\rho X_t} + \underset{(b)}{\sqrt{1-\rho^2}\,\eta_t} \tag{105}$$

which is an Ornstein-Uhlenbeck stochastic process in two parts, a deterministic part (a) and a random or stochastic part (b) expressing the uncertainty in the random process. X_{t+1} will have a normal distribution if both X_t and η_t are normally distributed because the central limit theorem states the sums of independent, normally distributed random variables are normally distributed. In the case of independence between successive X values, where $\rho = 0$, the deterministic part (a) is weighted by $\sqrt{0^2} = 0$ and the stochastic part (b) is weighted by $\sqrt{1^2} = 1$; so successive values of X are fully random. In the case of complete positive dependence between successive values of X, where $\rho = 1$, the deterministic part is fully in control, and each succeeding X_{t+1} is identical to its predecessor X_t.

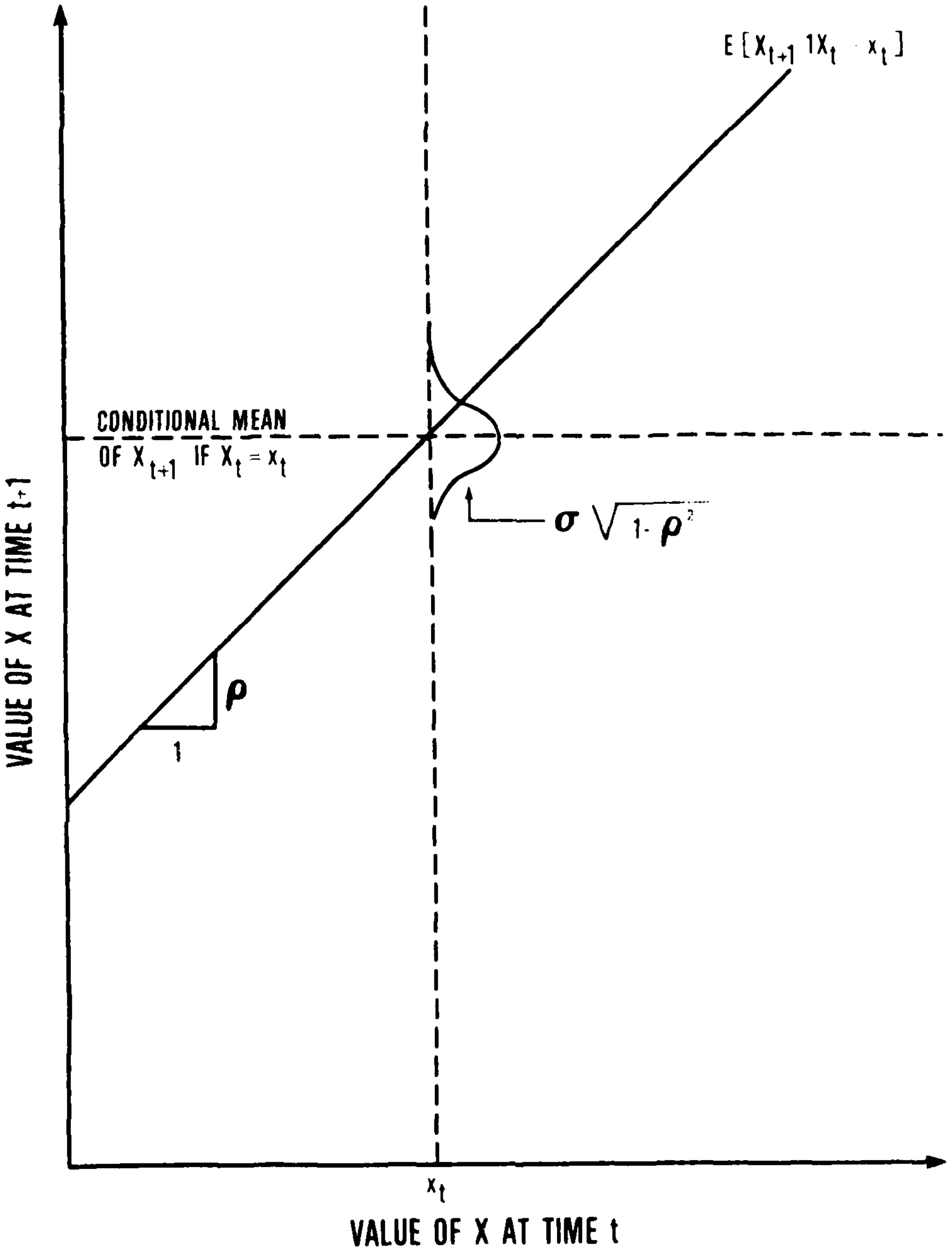


Figure 2. Conditional Distribution of X_{t+1} Given $X_t = x_t$.

Correlation in the intermediate case, where $0 < \rho < 1$, can be seen by recalling the definition of the Pearson product moment correlation coefficient (Equation 66)

$$r = \frac{\Sigma[(u-\bar{u})(v-\bar{v})]}{\sqrt{\Sigma(u-\bar{u})^2}\ \sqrt{\Sigma(v-\bar{v})^2}}$$

which can be rewritten as

$$r = \frac{1}{N}\ \frac{\Sigma(u-\bar{u})(v-\bar{v})}{s_u\ s_v} \tag{106}$$

$$= \frac{1}{N}\ \frac{\Sigma uv\ -\ \overline{uv}}{s_u\ s_v} \tag{107}$$

where the bars represent means or expected values, and s represents the standard deviation, the square root of the variance.

In applying Equation (107) to Equation (105) for $u = X_{t+1}$ and $v = X_t$, one finds that for standard normally distributed X,

$$\overline{X_t} = 0 \qquad \bar{X}_{t+1} = 0$$

$$s_{X_t} = 1 \qquad s_{X_{t+1}} = 1$$

and

$$r = \frac{1}{N}\ \Sigma\ X_{t+1}\ X_t \tag{108}$$

By substitution,

$$r = \rho\ \frac{\Sigma X_t{}^2}{N} + \sqrt{1-\rho^2}\ \frac{\Sigma\ (\eta_t\ X_t)}{N}$$

$$r = \rho E[X_t{}^2] + \sqrt{1-\rho^2}\ E[\eta_t X_t] \tag{109}$$

where E represents the expected value or mean. Since X_t is perfectly correlated with itself,

$$E[X_t{}^2] = 1$$

Furthermore, since η_t and X_t are independent of each other,

$$E[\eta_t X_t] = 0$$

and

$$r = \rho \tag{110}$$

for the Ornstein-Uhlenbeck model.

The question remains, does the time series model defined by Equation (104) reproduce a population with a specified mean and variance? The conditional mean of X_{t+1}, given that X_t equals x_t is

$$E[X_{t+1}|X_t=x_t] = E[\mu + \rho(x_t-\mu) + \sigma\sqrt{1-\rho^2}\ \eta_t] \qquad (111)$$

Since $E[\eta_t] = 0$, Equation 111 reduces to

$$E[X_{t+1}|x_t] = \mu + \rho(x_t-\mu)$$

which is the conditional mean specified by Equation (102). The conditional variance of X_{t+1} produced by Equation (104) is[7]

$$\begin{aligned} Var[X_{t+1}|x_t] &= E[\{X_{t+1} - E[X_{t+1}|x_t]\}^2|x_t] \\ &= E[\{\mu + \rho(x_t - \mu) + \sigma\sqrt{1-\rho^2}\ \eta_t - [\mu + \rho(x_t - \mu)]\}^2] \\ &= E[\sigma\sqrt{1-\rho^2}\ \eta_t]^2 \qquad (112) \end{aligned}$$

Since expected variance of η_t is equal to 1 $(E[Var(\eta_t)] = 1)$,

$$Var[X_{t+1}|x_t] = \sigma^2(1-\rho^2)$$

which is Equation (103). Thus the model produces distributions with the correct conditional mean and variance. The unconditional mean of X_{t+1} equals

$$E[X_{t+1}] = \mu + \rho\ (E[X_t] - \mu) + E[\eta_t]\sigma\ \sqrt{1-\rho^2} \qquad (113)$$

Once again noting that the mean of $\eta_t = 0$ and that the distribution of the variable is independent of time so that for all t, $E[X_{t+1}] = E[X_t] = E[X]$, it is clear that

$$(1-\rho)E[X] = (1-\rho)\mu \qquad (114)$$

or

$$E[X] = \mu \qquad (115)$$

The unconditional variance of the variable X is, using Equation (104),

[7] Note that the variance of X is given by $Var[X] = \sigma_X^2 = (1/N)\Sigma(X-\mu_X)^2 = E[(X - E[X])^2]$.

$$E[(X_{t+1} - \mu)^2] = E[\{\rho(X_t-\mu) + \sigma\sqrt{1-\rho^2}\, \eta_t\}^2]$$

$$= \rho^2 E[(X_t-\mu)^2] + 2\rho\sigma\sqrt{1-\rho^2}\, E[(X_t-\mu)\eta_t] + \sigma^2(1-\rho^2)\, E[\eta_t^2] \quad (116)$$

Since each value of η_t is independent, then $E[(X_t-\mu)\eta_t] = 0$ and $E[\eta_t{}^2] = 1$. Therefore, using the fact that $E[(X_{t+1}-\mu)^2] = E[(X_t-\mu)^2 = E[(X-\mu)^2]$, the unconditional variance of all X satisfies the equation,

$$(1-\rho^2)\, E[(X-\mu)^2] = (1-\rho^2)\sigma^2 \quad (117)$$

The conditional mean of X_{t+1} does not depend on the assumption that the random variables X_t and η_t are normally distributed. This relationship applies to all autoregressive Markov processes in the form of Equation (104), regardless of the distributions of X_t and η_t. However, if the variable X_t at time t is normally distributed with mean μ and variance σ^2 and if the η_t values are independently normally distributed with a mean of 0 and variance of 1, then the generated X's for $t \geq 1$ will also be normally distributed with mean μ and variance σ^2.

3.1.1 Transformation to the Normal Distribution. In order to apply Equation (105) to a weather variable that is not normally distributed, one must first transform the non-normal variable Z to its END $\overset{\shortparallel}{e}_Z$. The transformation to the normal distribution is referred to as transnormalization and is pictured graphically *in Figure* 3, which portrays the empirically determined cumulative frequency distribution of the ceiling at Scott AFB, IL, for February at 1200 LST, obtained from an historical weather tabulation called the Revised Uniform Summary of Surface Weather Observations (RUSSWO). Figure 3 actually shows an empirical estimate of the cumulative probability $Pr(C < c_T)$ of the cloud ceiling at Scott AFB, IL, 1200 LST, February, where C represents the ceiling in feet and c_T is some threshold value of the ceiling in feet. In the example shown, the probability that C is less than c_T = 5,000 ft is 0.365.

$$Pr(C < c_T) = 0.365$$

In the context of the normal probability distribution, this probability corresponds to some END $\overset{\shortparallel}{c}$. In other words, the integral of the standard normal density function $\phi(u)$ from $u = -\infty$ to $u = \overset{\shortparallel}{c}$ is $Pr(C < c_T)$, where

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) \quad (118)$$

and

$$\Phi(c_T) = Pr\{C < c_T\} = \int_{-\infty}^{\overset{\shortparallel}{c}} \phi(u)\, du \quad (119)$$

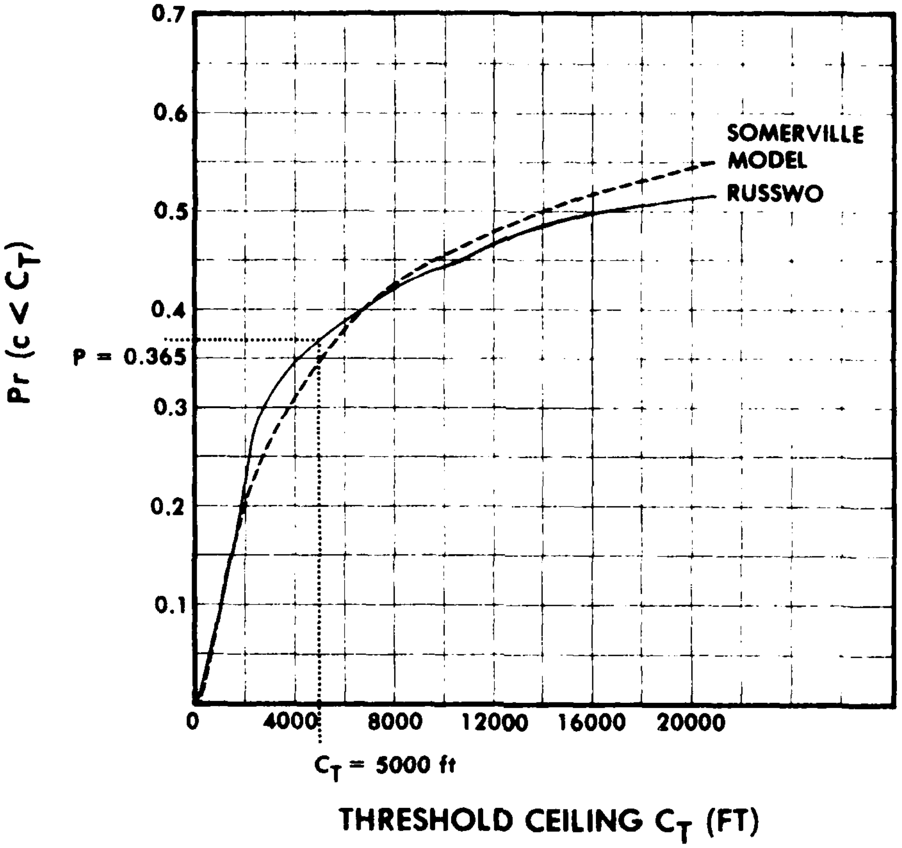


Figure 3. Cumulative Distribution Function of the Ceiling at Scott AFB IL, for February at 1200 LST. CDF extracted from the Scott RUSSWO and the model of Bean and Sommerville are shown.

The probability $Pr(C < c_T)$ is thus actually the area under the standard normal curve from $-\infty$ to $\ddot{c}$, as shown in Figure 4. Tables of integrals of the normal probability distribution (or rational approximations if one is working with a computer or calculator) show that a probability of 0.365 corresponds to an END, $\ddot{c} = -0.345$.

Transformation from the raw variable to its END can also be done graphically, using normal probability paper, as shown in Figure 5. The first step is to plot the cumulative distribution of the variable of interest, the ceiling in this case. Then one plots the cumulative normal distribution (a straight line on normal probability paper). One enters the graph with the raw variable (e.g., c_T = 5000 ft), proceeds vertically to the intersection with the observed distribution (e.g., at a probability of 0.365), then proceeds horizontally to the intersection with the cumulative normal distribution. From that intersection, one proceeds down and reads the value of the END (e.g., $\ddot{c} = -0.345$.)

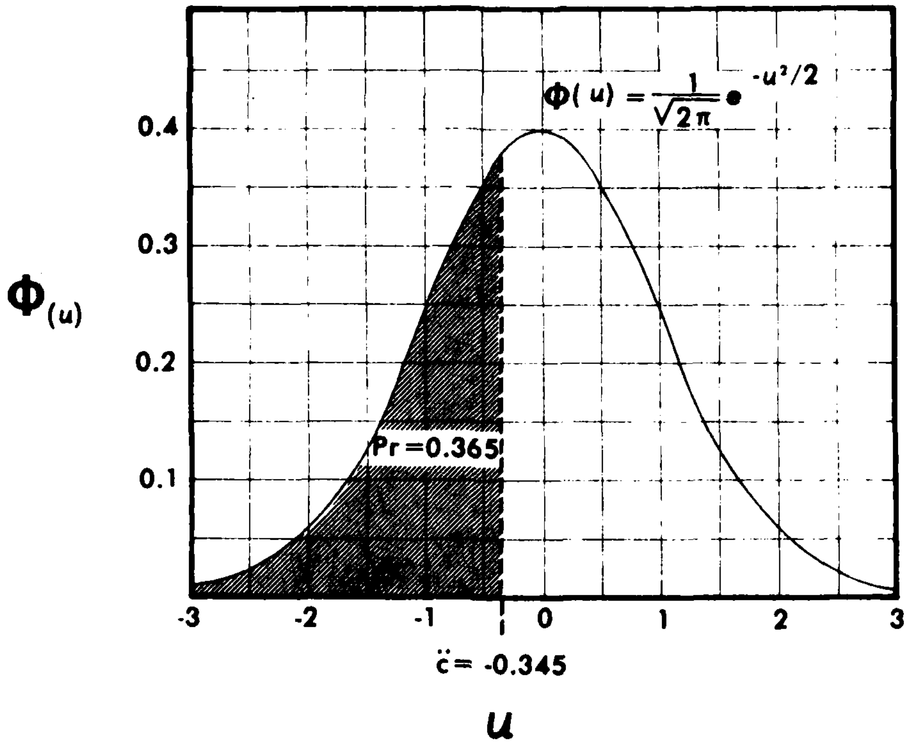


Figure 4. Normal Probability Distribution, Integrated from $-\infty$ to $\ddot{c}$ = -0.345, Yields a Cumulative Probability Pr = 0.365.

Thus, for 1200 LST in February at Scott, using RUSSWO data, a ceiling of 5000 ft corresponds to an END of -0.345. Because the RUSSWO is only an approximation to reality, it is better to say that 5000 ft corresponds approximately to an END of -0.345. A table of such approximate transformations is given below

Table 2. Transnormalization from Ceiling to END for Scott AFB, IL, February, 1200 LST.

Ceiling (ft)	Cumulative Probability	END
200	0.000	$-\infty$
1,000	0.104	-1.259
2,000	0.213	-0.796
3,000	0.305	-0.510
5,000	0.365	-0.345
10,000	0.440	-0.151
20,000	0.509	0.023

Using the normal transformation, then, every ceiling corresponds to an END of that ceiling. Since ENDs are in themselves normally distributed with a mean of

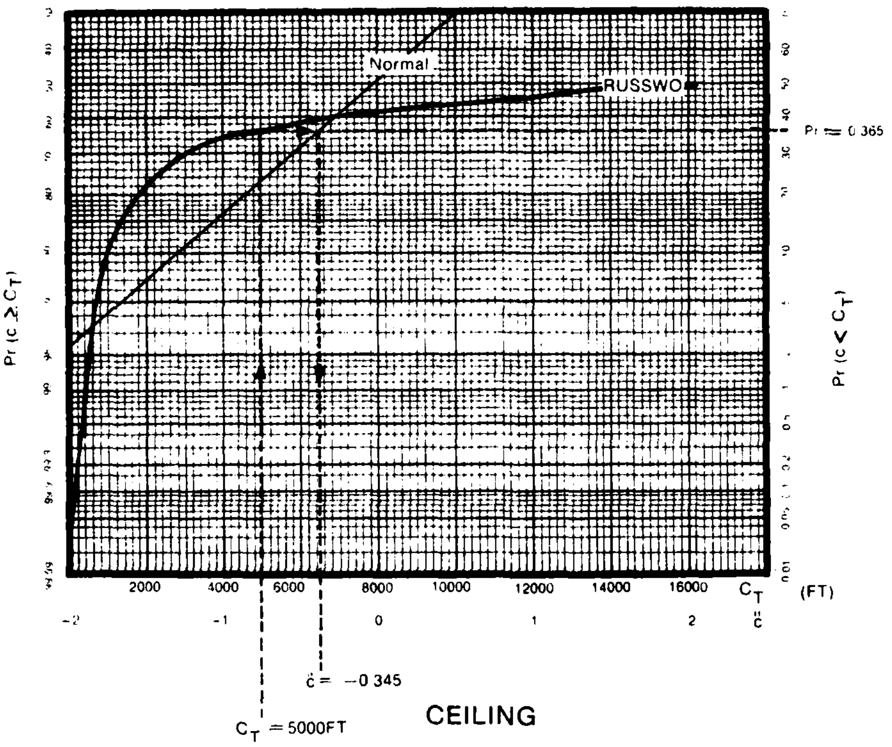


Figure 5. Cumulative Distribution Function of the Ceiling at Scott AFB IL, for February at 1200 LST, as Extracted from the Scott RUSSWO, Plotted on Normal Probability Paper. The straight line is the cumulative normal distribution.

zero and variance of one, they can be used as random variables in the Ornstein-Uhlenbeck process, Equation (105). Such a process, applied to the ceiling (not normally distributed in general) is

$$\ddot{c}_{t+1} = \rho_{cc}\,\ddot{c}_t + \sqrt{1-\rho_{cc}^2}\;\eta_t \qquad (120)$$

where $\ddot{c}$ values are ENDs of the ceiling C.

3.1.2 Simulation of the Cloud Ceiling. To see how such a simulation might work in practice, consider a case with an initial ceiling at 2000 ft (the corresponding END $\ddot{c}$ is -0.796). Assume a correlation according to Gringorten's model,

$$\rho_{cc} = 0.945^{\Delta t} \qquad (121)$$

where Δt is the time step, unity in this case. We generate a random normal number, for example $\eta_t = 0.325$. Applying the Ornstein-Uhlenbeck process in Equation (120) yields

$$\begin{aligned}\overset{''}{c}_{t+1} &= (0.945)(-0.796) + (\sqrt{1-0.945^2})\,(0.325) \\ &= (0.945)(-0.796) + (0.327)(0.325) \\ &= -0.752 + 0.106 \\ \overset{''}{c}_{t+1} &= -0.646\end{aligned}$$

which corresponds to a ceiling of about 2200 ft. At the next time step, $\overset{''}{c}_t$ becomes -0.646. Another random normal number is drawn, say -0.102. Then

$$\begin{aligned}\overset{''}{c}_{t+1} &= (0.945)(-0.646) + (0.327)(-0.102) \\ &= -0.610 + 0.033 \\ \overset{''}{c}_{t+1} &= -0.643\end{aligned}$$

which again corresponds to a ceiling of about 2200 ft.

If continued, this process will generate a time series of the ceiling whose probability distribution is the same as the distribution specified initially (e.g., Figure 3), within the limits imposed by sampling error. The process will not necessarily produce the same durations as those of the original data. The distribution of durations of, for example, low ceiling episodes is affected by the parameter ρ_{cc} and by the first order Markov assumption. It is possible to determine a value of ρ_{cc} that will best "fit" a given distribution of durations.

3.2 Two-variable, Single-station Model (V2S1)

The simulation model expressed in Equation (105) is severely limited, in the sense that it can be applied only to a time series of a single variable, such as ceiling or sky cover. One is frequently interested in simulating more than one variable (e.g., ceiling and visibility) in such a manner as to preserve the cross-correlations between them. The V2S1 model handles the two variable case by including two time series of ENDs, one END for each of the two variables, and then carrying the cross-correlation information in the stochastic part of the solution. For example, there is an END for the ceiling, $\overset{''}{c}$, and an END for the visibility, $\overset{''}{v}$. These advance by separate Ornstein-Uhlenbeck equations

As in Equation (120) $$\overset{''}{c}_{t+1} = \rho_{cc}\,\overset{''}{c}_t + \sqrt{1-\rho_{cc}{}^2}\;\eta_c$$

$$\overset{''}{v}_{t+1} = \rho_{vv}\,\overset{''}{v}_t + \sqrt{1-\rho_{vv}{}^2}\;\eta_v \qquad (122)$$

But because it is desired to produce time series of ceiling and visibility that are correlated across variables (i.e., cross-correlated), the stochastic parts of Equations (120) and (122) must be linked. This is done by generating a random normal number of visibility η_v that is correlated with that, η_c, previously generated for ceiling. To do this, the procedure is first to generate an independent η_c and then to set

$$\eta_v = \rho'_{cv}\,\eta_c + \sqrt{1-\rho'^{2}_{cv}}\,\eta \qquad (123)$$

where η is another independent random normal number, and ρ'_{cv} is proportional to the cross-correlation between ENDs of ceiling and visibility. Equation (123) is essentially the generation algorithm for producing ENDs having the correlation ρ'_{cv}.

In the case of independence when $\rho'_{cv} = 0$, $\eta_v = \eta$ and Equations (120) and (122) generate unrelated time series of ceiling and visibility. In the case of perfect positive correlation, when $\rho'_{cv} = 1$,

$$\eta_v = \eta_c$$

the time series for visibility will depend completely on that for the ceiling. Indeed, if ρ_{cc} and $\rho_{vv} = 1$, the two time series will be identical except for a shift due to differing initial values. In the intermediate case, ceiling and visibility will be partially correlated according to the value of ρ'_{cv}, which is proportional to the correlation ρ_{cv} between ceiling and visibility. The process is depicted in Figure 6.

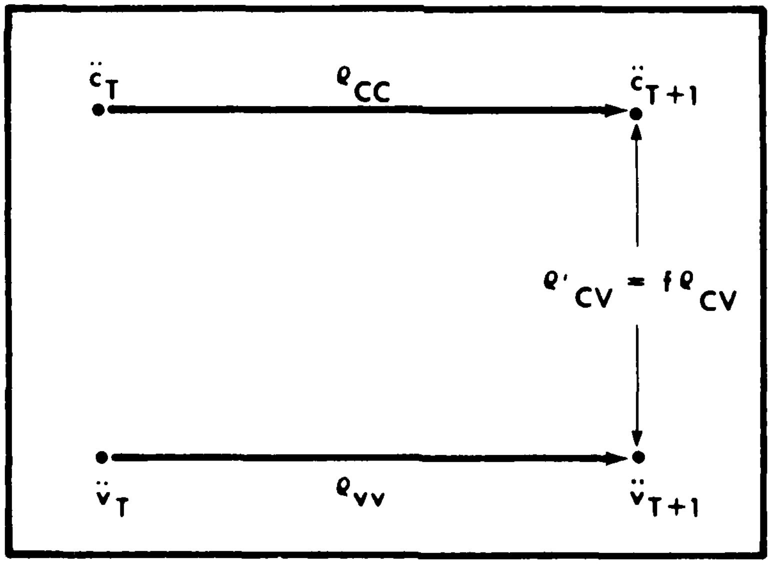


Figure 6. The Weather-A Process.

The serial correlation ρ_{cc} between ceiling at t and ceiling at t+1 is preserved, as is the correlation ρ_{vv} between visibility at t and visibility at t+1. The correlation ρ_{cv} between ceiling and visibility at the same time is proportional to ρ'_{cv} through a constant of proportionality f.

It is instructive to consider how the cross-correlation ρ_{cv} between ceiling and visibility relates to ρ'_{cv}. Using Equations (120), (122), and (123) in Equation (108) produces the equation,

$$\rho_{cv} = \frac{1}{N} \Sigma \; \{[\rho_{cc} \overset{''}{c}_t + \sqrt{1-\rho_{cc}{}^2}\, \eta_c] \cdot [\rho_{vv} \overset{''}{v}_t + \sqrt{1-\rho_{cc}{}^2}\, (\rho'_{cv}\, \eta_c + \sqrt{1-\rho'_{cv}{}^2}\, \eta)]\} \qquad (124)$$

which expands to

$$\begin{aligned}\rho_{cv} &= \Sigma\, \rho_{cc}\, \rho_{cv}\, \overset{''}{c}_t\, \overset{''}{v}_t \\ &+ \frac{1}{N} \Sigma\, \rho_{cc}\, \rho'_{cv}\, \sqrt{1-\rho_{vv}{}^2}\, \eta_c \overset{''}{c}_t \\ &+ \frac{1}{N} \Sigma\, \rho_{cc}\, \sqrt{1-\rho_{vv}{}^2}\, \sqrt{1-\rho'_{cv}{}^2}\, \eta \overset{''}{c}_t \\ &+ \frac{1}{N} \Sigma\, \rho_{vv}\, \sqrt{1-\rho_{cc}{}^2}\, \eta_c \overset{''}{v}_t \\ &+ \frac{1}{N} \Sigma\, \rho'_{cv}\, \sqrt{1-\rho_{cc}{}^2}\, \sqrt{1-\rho_{vv}{}^2}\, \eta_c{}^2 \\ &+ \frac{1}{N} \Sigma\, \sqrt{1-\rho_{cc}{}^2}\, \sqrt{1-\rho_{vv}{}^2}\, \sqrt{1-\rho'_{cv}{}^2}\, \eta_c \eta \qquad (125)\end{aligned}$$

Because $E[\eta_c \overset{''}{c}_t] = E[\eta \overset{''}{c}_t] = E[\eta_c \overset{''}{v}_t] = E[\eta_c \eta] = 0$, due to independence, and because $E[\eta_c{}^2] = 1$ due to dependence,

$$\rho_{cv} = \rho_{cc}\, \rho_{vv}\, \frac{\Sigma\, \overset{''}{c}_t\, \overset{''}{v}_t}{N} + \rho'_{cv}\, \sqrt{1-\rho_{cc}{}^2}\, \sqrt{1-\rho_{vv}{}^2} \qquad (126)$$

But

$$\frac{\Sigma\, \overset{''}{c}_t\, \overset{''}{v}_t}{N} = \rho_{cv}$$

so

$$\rho_{cv} = \rho_{cc}\, \rho_{vv}\, \rho_{cv} + \rho'_{cv}\, \sqrt{1-\rho_{cc}{}^2}\, \sqrt{1-\rho_{vv}{}^2} \qquad (127)$$

or

$$\rho_{cv} = \frac{\sqrt{1-\rho_{cc}^2}\ \sqrt{1-\rho_{vv}^2}}{1 - \rho_{cc}\ \rho_{vv}}\ \rho'_{cv} \qquad (128)$$

From Equation (128) it can be seen that to obtain a correlation ρ_{cv} between ceiling and visibility, it is necessary to use a model correlation parameter ρ'_{cv} given by

$$\rho'_{cv} = \frac{1 - \rho_{cc}\ \rho_{vv}}{\sqrt{1-\rho_{cc}^2}\ \sqrt{1-\rho_{vv}^2}}\ \rho_{cv} = f\rho_{cv} \qquad (129)$$

The factor f reduces to 1 when $\rho_{cc} = \rho_{vv}$ but otherwise is greater than 1, so

$$\rho'_{cv} \geq \rho_{cv}$$

This is illustrated in Figure 7, in which it is desired to obtain $\rho_{cv} = 0.3$. The ρ'_{cv} necessary to obtain that ρ_{cv} value is 0.3 if $\rho_{cc} = \rho_{cv}$. Where $\rho_{cc} \neq \rho_{vv}$, the ρ'_{cv} needed to obtain $\rho_{cv} = 0.3$ is greater than 0.3. For example, if $\rho_{cc} = 0.8$ and $\rho_{vv} = 0.4$ then $\rho_1 = 0.8$, $\rho_2 = 0.4$, $\rho'_{cv} = 0.37$. The ratio $f = \rho'_{cv}/\rho_{cv}$ can be quite large for cases where ρ_{cc} and ρ_{vv} differ substantially. Figure 8 gives the factor f as a function of ρ_{cc} and ρ_{cv}.

It can be seen from Equation (123) that real solutions can be obtained only if ρ'_{cv} is less or equal to unity. Hence,

$$f\ \rho_{cv} \leq 1 \qquad (130)$$

$$\rho_{cv} \leq \frac{1}{f}$$

Thus, the mathematics imposes an upper limit on the cross-correlation this model is capable of producing between ceiling and visibility. For the example given above, in which $\rho_{cc} = 0.8$ and $\rho_{cv} = 0.4$, f = 1.24 and ρ_{cv} cannot exceed 0.81. In this case, the model in its present form cannot simulate phenomena "c" and "v" whose ENDs are cross-correlated more strongly than 0.81. This upper limit on ρ_{cv} depends on ρ_{cc} and ρ_{vv} and must be treated on a case-by-case basis. In the special case where $\rho_{cv} = \rho'_{cv}$, cross-correlation values up to 1.0 can be simulated.

The V2S1 model does not explicitly preserve what is known as the cross-lag correlation, such as $\rho_{v_t c_{t+1}}$, the correlation between the visibility at time t and the ceiling at time t+1. This can seen by applying Equation (107) to Equations (120) and (122).

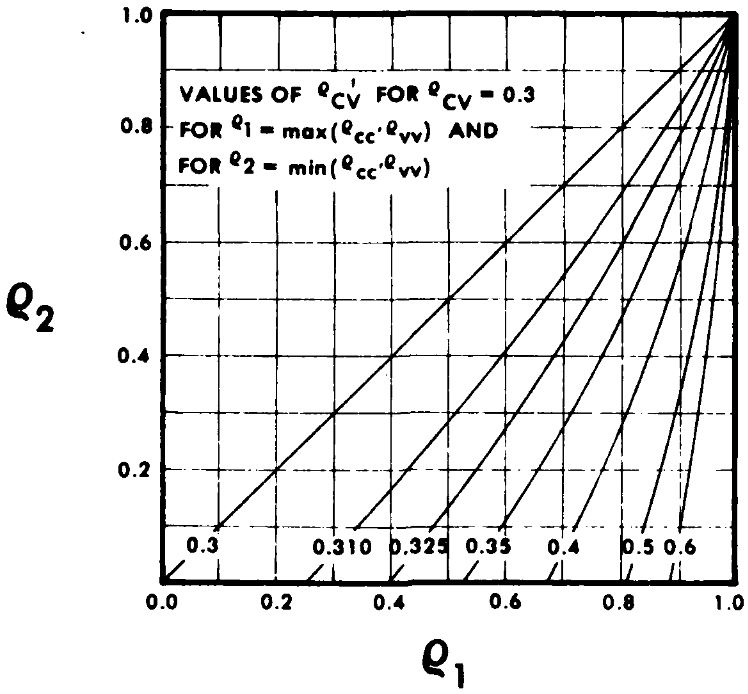


Figure 7. Values of ρ_{cv}' for $\rho_{cv} = 0.3$.

$$\rho_{c_{t+1} v_t} = \frac{1}{N} \Sigma \ddot{c}_{t+1} \ddot{v}_t$$

$$= \frac{1}{N} \Sigma (\rho_{cc} \ddot{c}_t + \sqrt{1-\rho_{cc}^2}\ \eta_c) \ddot{v}_t$$

$$= \rho_{cc} \frac{\Sigma \ddot{c}_t \ddot{v}_t}{N} + \sqrt{1-\rho_{cc}^2}\ \frac{\Sigma \ddot{v}_t \eta_c}{N} \qquad (131)$$

Because $\ddot{v}_t$ and η_c are independent, the final term is zero, and

$$\rho_{c_{t+1} v_t} = \rho_{cc}\ \rho_{cv} \qquad (132)$$

In other words, in the V2S1 model, the cross-lag correlation reduces to the product of autocorrelation of the ceiling and the cross-correlation of the ceiling and visibility. This is equivalent to saying that in this model the cross-lag correlation reduces to the automatic correlation between ceiling and visibility

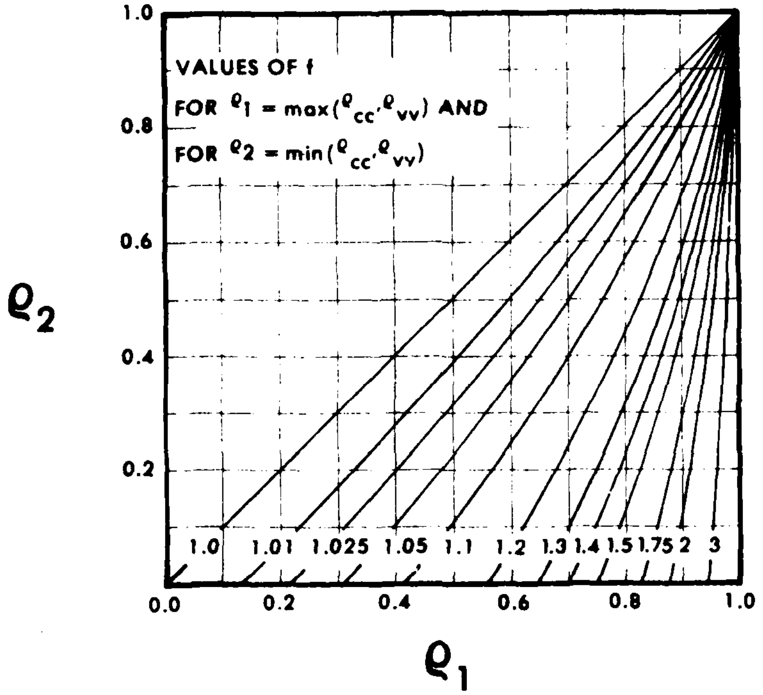


Figure 8. Values of f.

Panofsky and Brier (1968) describe automatic correlation in terms of two variables p and q that are each separately correlated with a third variable s. The cross-correlations are

$$\rho_{ps} \text{ and } \rho_{sq}$$

The separate correlations between s and p and between s and q guarantee an "automatic" correlation between p and q even if the two are not intrinsically related. The automatic correlation would be the product $\rho_{ps}\ \rho_{sq}$.

Application to the V2S1 model is shown in Figure 9, a correlation influence diagram. $\ddot{c}_t$ is correlated with $\ddot{c}_{t+1}$ by autocorrelation ρ_{cc}. $\ddot{c}_t$ is correlated with $\ddot{v}_t$ by cross-correlation $\ddot{c}$ ρ_{cv} . This guarantees an automatic correlation of $\rho_{cc} \cdot \rho_{cv}$ between $\ddot{v}_t$ and $\ddot{c}_{t+1}$. The model cross-lag correlation given by Equation (132) is the automatic correlation. Hence, in the V2S1 model the cross-lag correlation reduces to the automatic correlation.

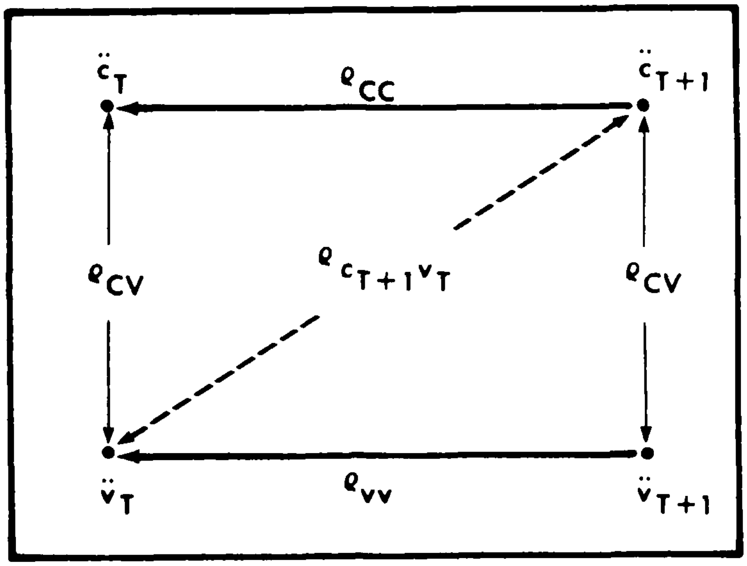


Figure 9. Correlation Influence Diagram for the Weather-A Process. The cross-lag correlation is shown as a dotted line because it reduces to automatic correlation in this model.

Whether this true in Nature is another question. A model is a simplification or generalization of Nature. Work conducted to date gives no indication that reducing the cross-lag correlation to the automatic correlation has any adverse affect on the model as a weather simulator. The model's originator believes that cross-lag correlations between ceiling and visibility are very nearly equal to automatic correlation. Whether this is true or at least approximately true for other variables is subject to verification using actual data.

3.3 Modeling Cumulative Distribution Functions

3.3.1 Graphical Approach. The V2S1 model produces correlated time series of ENDs such as $\ddot{c}$ and $\ddot{v}$. These can be translated into raw variables by a graphical inverse transnormalization procedure such as that described previously for Scott AFB in February. The graphical approach is limited and cannot be applied in the computer.

3.3.2 Tabular Approach. The other procedure described previously involved using the graphical approach such as in Figure 3 to construct tables, such as Table 2, relating ceiling heights to their ENDs. This can be accomplished using any cumulative distribution found in RUSSWOs.

There are several important problems with this approach. First, any table is discrete. The model generates continuous END values such as $\overset{''}{c} = -0.441$. Table 2 contains no such value. Interpolation, which introduces error, would be required to translate such an END into its corresponding ceiling height. Secondly, RUSSWOs contain errors and biases introduced by the weather observing system. These often show up as bumps or spikes in the relative frequencies, corresponding to reportable values, popular values, location of visibility markers, etc. Finally, from the point of view of simulation, it is inefficient to maintain in computer storage entire RUSSWOs from which to interpolate probabilities.

3.3.3 Distribution Fitting Approach. An increasingly popular alternative to storing RUSSWOs is to *model* the RUSSWO probabilities, using regression techniques to fit variously shaped probability distributions or curves to RUSSWO data. The result of this is a continuous function of the form,

$$P = F(x) \qquad (133)$$

from which continuous probability estimates can be obtained simply by evaluating the function. Correspondingly, continuous variable estimates can be obtained by evaluating the function inverse

$$x = F^{-1}(P) \qquad (134)$$

Considerable work of this sort has been done by Somerville and Bean (see references) for ceiling, visibility, sky cover, and rainfall. See Appendix A for a list of the functions that USAFETAC uses to model cumulative frequency distributions of various meteorological variables.

For example, Bean fitted the three-parameter Burr curve to cumulative distributions of ceiling,

$$Pr(X < x_T) = 1 - [1 + (\frac{x_t}{c})^a]^{-b} \qquad (135)$$

Somerville fitted the two-parameter Weibull distribution to the cumulative distributions of visibility,

$$Pr(V < v_T) = 1 - \exp(-\alpha v_T{}^{\beta}) \qquad (136)$$

A comparison between the Scott AFB RUSSWO and the Burr distribution fit for the ceiling at 1200 LST in February can be seen in Figure 3.

O'Connor of USAFETAC has applied log cubic and inverse linear equations to the ceiling and visibility (see Friend, 1978). Of these, the log cubic, namely,

$$\Pr(X < x_T) = c_1 + c_2 \ln x_T + c_3(\ln x_T)^2 + c_4(\ln x_T)^3 \qquad (137)$$

has been applied to both ceiling and visibility. Log cubics have been fitted to the cumulative distribution functions of both ceiling and visibility by two least squares linear regression methods

- Inverting the normal equations by Gaussian elimination using the subroutines DECOMP and SOLVE of Forsythe, Malcolm, and Moler (1977).

- Singular value decomposition, using the subroutine SVD of Forsythe, Malcolm, and Moler (1977).

The inverse linear curve of O'Connor, namely,

$$\Pr(X > x_T) = \frac{1}{fx_T + g} \qquad (138)$$

has been fitted to cumulative probability distributions of visibility by use of DECOMP and SOLVE.

Comparisons between the various curve fits for ceiling and visibility in winter and summer at Scott AFB, IL, and Kitzingen AAF, Germany (EDIN, WMO 106590, 47°45'N, 10°13'E), are shown in Tables 3-14. In these tables it can be seen that some of the fits were done over the entire range of the variable whose cumulative distribution was being modeled, and other fits were done over a restricted range. The curve fits are evaluated in terms of a root mean squared difference (RMS) between the RUSSWO value and the modeled value (see Appendix B for an explanation of the RMS equation). Where appropriate, an additional evaluation is provided, limited to a portion of the total range of the variable whose cumulative distribution function was modeled.

Taken as a whole, the curve fit results for ceiling show the clear superiority of Bean and Somerville's Burr curve. The log cubic as fitted by SVD is at times competitive, especially if one's concern is only with ceilings of 10,000 ft or less. In many cases, by extending the curve fit to 20,000 ft, where data are relatively unreliable, the fit for the portion of the curve below 10,000 ft is impaired. The curve fits for Kitzingen were noticeably poorer than those for Scott, due to the differences in the shapes of the curves for the two stations, as seen in Figure 10.

Table 3. Curve Fit Information for Ceiling Data at Scott AFB, IL, January, 0600 LST.

Threshold Ceiling (ft)	RUSSWO (PCT)	Bean and Somerville Burr Curve to 10,000 ft (PCT)	Bean and Somerville Burr Curve to 20,000 ft (PCT)	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft (PCT)	O'Connor Log Cubic by SVD to 20,000 ft (PCT)	O'Connor Log Cubic by SVD to 10,000 ft (PCT)
20,000	48.8	45.0	45.7	40.2	46.8	42.3
10,000	55.6	53.4	54.1	52.4	55.1	53.0
3,000	67.7	70.0	70.6	70.4	69.8	70.1
2,000	73.9	75.7	76.2	75.7	74.8	75.3
1,000	85.1	84.3	84.7	83.8	82.9	83.6
200	98.3	96.0	96.2	98.7	99.5	99.2
0	100.0	100.0	100.0	100.0	99.6	99.7
RMS:						
(Eval to 10,000 ft)		1.8	1.8	1.8	1.4	1.7
(Eval to 20,000 ft)		2.2	2.0	3.7	1.5	2.9

Table 4. Curve Fit Information for Ceiling Data at Scott AFB, IL, February, 1200 LST.

Threshold Ceiling (ft)	RUSSWO (PCT)	Bean and Somerville Burr Curve to 10,000 ft (PCT)	Bean and Somerville Burr Curve to 20,000 ft (PCT)	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft (PCT)	O'Connor Log Cubic by SVD to 20,000 ft (PCT)	O'Connor Log Cubic by SVD to 10,000 ft (PCT)
20,000	49.1	45.2	45.8	35.8	45.3	36.9
10,000	56.0	54.1	54.7	51.6	55.3	52.1
3,000	69.5	72.5	73.3	73.7	73.0	73.5
2,000	78.7	79.0	79.8	79.7	78.3	79.4
1,000	89.6	88.7	89.5	88.3	86.9	88.1
200	100.0	98.6	98.9	101.3	103.0	101.8
0	100.0	100.0	100.0	100.0	99.4	99.6
RMS:						
(Eval to 10,000 ft)		1.6	1.8	2.6	2.2	2.5
(Eval to 20,000 ft)		2.1	2.1	5.6	2.5	5.2

Table 5. Curve Fit Information for Ceiling Data at Scott AFB, IL, July, 1800 LST.

Threshold Ceiling	RUSSWO	Bean and Somerville Burr Curve to 10,000 ft	Bean and Somerville Burr Curve to 20,000 ft	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft	O'Connor Log Cubic by SVD to 20,000 ft	O'Connor Log Cubic by SVD to 10,000 ft
(ft)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)
20,000	74.4	64.6	73.0	77.3	72.9	75.4
10,000	84.8	83.1	83.9	86.0	84.3	85.5
3,000	97.7	97.3	97.2	96.0	96.4	96.3
2,000	99.3	99.8	98.8	98.1	100.8	98.4
1,000	99.9	100.0	99.7	100.4	101.2	100.4
200	100.0	100.0	100.0	100.0	99.7	99.7
0	100.0	100.0	100.0	100.0	100.1	100.1
RMS:						
(Eval to 10,000 ft)		0.7	0.5	1.0	0.9	0.8
(Eval to 20,000 ft)		3.8	0.7	1.4	1.0	0.8

Table 6. Curve Fit Information for Ceiling Data at Kitzingen AAF, Germany, January, 0600 LST.

Threshold Ceiling	RUSSWO	Bean and Somerville Burr Curve to 10,000 ft	Bean and Somerville Burr Curve to 20,000 ft	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft	O'Connor Log Cubic by SVD to 20,000 ft	O'Connor Log Cubic by SVD to 10,000 ft
(ft)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)
20,000	26.9	19.3	20.2	-9.7	15.3	-9.5
10,000	27.3	28.2	30.5	19.8	31.1	20.1
3,000	53.5	54.2	58.1	59.2	57.6	59.1
2,000	69.0	66.7	68.7	69.3	66.0	69.1
1,000	87.0	87.7	84.3	83.2	79.6	82.9
200	98.1	99.8	98.3	100.9	105.8	101.5
0	100.0	100.0	100.0	100.0	98.8	99.5
RMS:						
(Eval to 10,000 ft)		1.0	2.5	4.3	5.1	4.3
(Eval to 20,000 ft)		3.0	3.5	14.4	6.4	14.3

Table 7. Curve Fit Information for Ceiling Data at Kitzingen AAF, Germany, February, 1200 LST.

Threshold Ceiling	RUSSWO	Bean and Somerville Burr Curve to 10,000 ft	Bean and Somerville Burr Curve to 20,000 ft	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft	O'Connor Log Cubic by SVD to 20,000 ft	O'Connor Log Cubic by SVD to 10,000 ft
(ft)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)
20,000	33.2	28.3	29.9	-2.7	22.3	-4.0
10,000	35.6	37.9	39.5	28.1	40.3	28.0
3,000	61.1	62.9	64.2	67.6	66.9	67.8
2,000	80.9	74.5	75.4	77.3	74.6	77.3
1,000	93.4	94.3	94.3	90.0	86.3	89.6
200	99.6	100.0	100.0	102.5	106.0	102.9
0	100.0	100.0	100.0	35.2	98.8	99.6
RMS:						
(Eval to 10,000 ft)		2.9	3.1	26.9	5.6	4.8
(Eval to 20,000 ft)		3.3	3.1	28.3	6.6	14.8

Table 8. Curve Fit Information for Ceiling Data at Kitzingen AAF, Germany, July, 1800 LST.

Threshold Ceiling	RUSSWO	Bean and Somerville Burr Curve to 10,000 ft	Bean and Somerville Burr Curve to 20,000 ft	O'Connor Log Cubic by DECOMP & SOLVE to 10,000 ft	O'Connor Log Cubic by SVD to 20,000 ft	O'Connor Log Cubic by SVD to 10,000 ft
(ft)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)	(PCT)
20,000	57.0	38.4	53.0	42.6	48.0	37.1
10,000	60.5	60.5	65.5	64.4	68.1	62.9
3,000	93.9	94.1	93.1	89.9	90.5	90.8
2,000	98.0	97.7	98.7	95.3	95.0	96.0
1,000	99.5	99.6	100.0	101.1	99.9	101.1
200	100.0	100.0	100.0	99.9	100.9	99.1
0	100.0	100.0	100.0	100.0	99.9	100.3
RMS:						
(Eval to 10,000 ft)		0.2	2.1	2.6	3.6	1.9
(Eval to 20,000 ft)		7.0	2.5	6.0	4.8	7.7

Table 9. Curve Fit Information for Visibility Data at Scott AFB, IL, January, 0600 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	51.6	49.1	54.8	55.0
4.0	68.2	65.4	65.2	66.8
3.0	75.5	74.6	72.0	73.9
2.0	83.7	83.9	80.4	82.1
1.0	91.8	93.0	91.0	91.7
0.5	96.6	97.1	97.5	96.8
0.0	100.0	100.0	104.8	100.1
RMS		1.5	3.1	1.6

Table 10. Curve Fit Information for Visibility Data at Scott AFB, IL, February, 1200 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	69.5	68.3	73.1	63.9
4.0	83.6	82.6	81.1	76.7
3.0	88.8	89.0	85.8	83.9
2.0	92.8	94.4	91.1	91.8
1.0	98.0	98.2	97.0	99.7
0.5	99.8	99.5	100.3	101.7
0.0	100.0	100.0	103.8	100.3
RMS		0.8	2.6	4.0

Table 11. Curve Fit Information for Visibility Data at Scott AFB, IL, July, 1800 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	94.9	94.6	96.3	96.6
4.0	98.6	98.8	97.7	98.0
3.0	99.4	99.6	98.5	98.7
2.0	99.8	99.9	99.2	99.5
1.0	100.0	100.0	99.9	100.2
0.5	100.0	100.0	100.3	100.3
0.0	100.0	100.0	100.7	100.0
RMS		0.2	0.8	0.8

Table 12. Curve Fit Information for Visibility Data at Kitzingen AAF, Germany, January, 0600 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	27.1	25.8	31.0	26.4
4.0	46.2	48.9	41.9	48.5
3.0	63.7	63.6	50.8	61.4
2.0	81.0	78.9	64.7	75.9
1.0	92.2	92.4	88.7	92.1
0.5	96.9	97.4	108.9	99.1
0.0	100.0	100.0	141.0	100.2
RMS		1.4	18.2	2.4

Table 13. Curve Fit Information for Visibility Data at Kitzingen AAF, Germany, February, 1200 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	54.0	53.2	57.0	55.7
4.0	69.2	70.2	67.7	69.4
3.0	78.3	79.1	74.8	77.4
2.0	87.6	87.7	83.5	86.4
1.0	97.1	95.2	94.5	96.4
0.5	99.4	98.2	101.1	100.6
0.0	100.0	100.0	108.7	100.1
RMS		1.0	4.2	1.0

Table 14. Curve Fit Information for Visibility Data at Kitzingen AAF, Germany, July, 1800 LST.

Threshold Visibility	RUSSWO	Somerville and Bean Weibull Curve	O'Connor Inverse Linear by DECOMP & SOLVE	O'Connor Log Cubic by SVD
(SM)	(PCT)	(PCT)	(PCT)	(PCT)
6.0	96.2	96.7	96.9	97.0
4.0	98.3	98.5	98.0	98.1
3.0	99.1	99.1	98.6	98.8
2.0	99.8	99.6	99.2	99.4
1.0	99.8	99.9	99.9	100.1
0.5	100.0	100.0	100.2	100.2
0.0	100.0	100.0	100.5	100.0
RMS		0.2	0.5	0.4

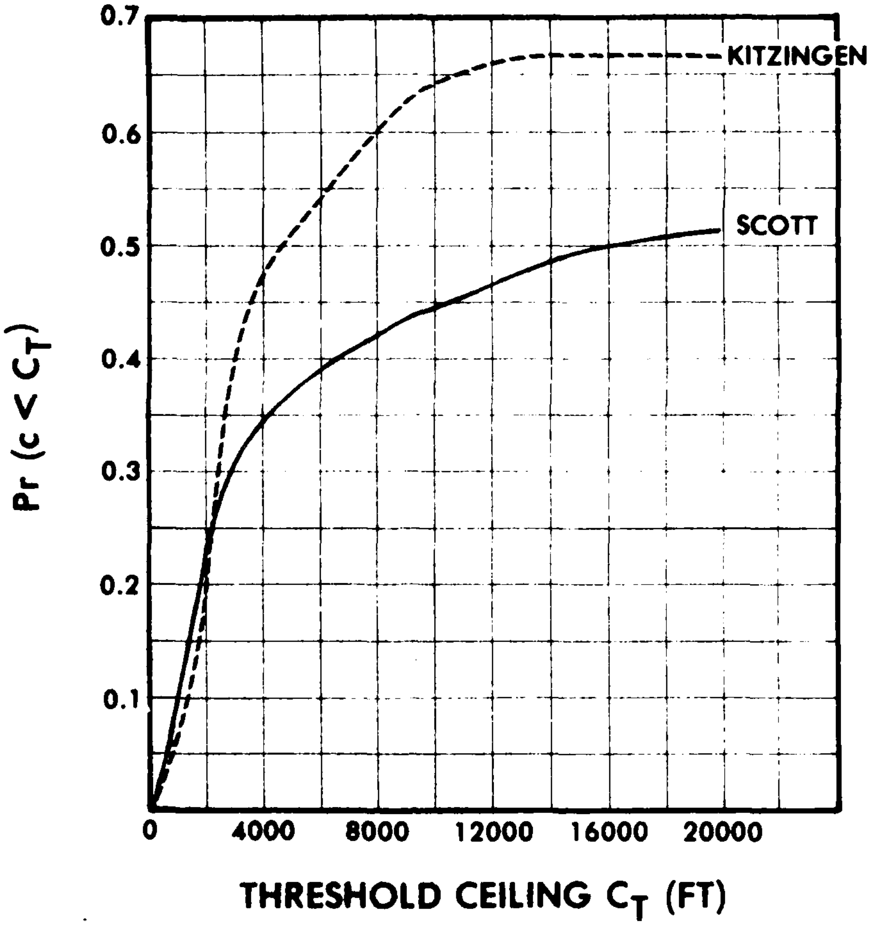


Figure 10. Cumulative Distribution Functions (CDF) of the Ceiling at Scott AFB IL, and Kitzingen AAF Germany, for February at 1200 LST, Extracted from the RUSSWOs.

The curve fit results for visibility again demonstrate the superiority of Somerville and Bean's function. The inverse linear function behaves somewhat unreliably; it does well sometimes and poorly at other times. The greater reliability of the log cubic makes it second best after Somerville and Bean's Weibull curve.

3.4 Tests of the V2S1 Model

3.4.1 Correlation. Lengthy runs of the V2S1 model were made to test whether the correlation behavior predicted by Equations (110) and (129) were in fact shown. Results for runs of 10,000 and 100,000 simulated hourly observations are shown in Table 15. Bracketing values in parentheses represent the 95-percent confidence limits for each correlation coefficient, determined according to the procedures in the following paragraphs.

Table 15. Correlation Tests.

PARAMETERS SPECIFIED TO MODEL			PREDICTED VALUE OF	PARAMETERS RECOVERED FROM MODEL				
ρ_{cc}	ρ_{vv}	ρ_{cc}	$\rho_{v_t c_{t+1}}$	r_{cc}	r_{vv}	r_{cv}	$r_{v_t c_{t+1}}$	N
(0.931)	(0.931)	(0.275)	(0.260)					
0.945	0.945	0.300	0.284	0.946	0.944	0.300	0.284	10000
(0.956)	(0.956)	(0.324)	(0.308)					
(0.931)	(0.824)	(0.275)	(0.260)					
0.945	0.845	0.300	0.284	0.946	0.837	0.294	0.278	10000
(0.956)	(0.863)	(0.324)	(0.308)					
(0.778)	(0.177)	(0.275)	(0.216)					
0.800	0.200	0.300	0.240	0.801	0.184	0.289	0.227	10000
(0.820)	(0.223)	(0.324)	(0.263)					
(0.778)	(0.177)	(0.474)	(0.375)					
0.800	0.200	0.500	0.400	0.801	0.186	0.494	0.391	10000
(0.820)	(0.223)	(0.525)	(0.424)					
(0.793)	(0.793)	(0.292)	(0.233)					
0.800	0.800	0.300	0.240	0.797	0.800	0.300	0.240	100000
(0.807)	(0.807)	(0.308)	(0.247)					
(0.793)	(0.193)	(0.292)	(0.233)					
0.800	0.200	0.300	0.240	0.797	0.198	0.300	0.239	100000
(0.807)	(0.207)	(0.308)	(0.247)					

Correlation coefficients r that are calculated from sample data (whether historical data or data generated by a simulation model such as V2S1) are subject to sampling variability. The distribution of the sample correlation coefficients r is not normal but approaches normality as the sample size increases. The approach of the distribution of r to normality depends not only on sample size but also on the value of the population correlation ρ. If samples are drawn from a population for which $\rho = 0$, the distribution is approximately normal, approaching normality rather slowly as the sample size increases. In this case, Student's t-distribution or the normal distribution is used in testing inferences about ρ. If samples are drawn from a population for which $\rho \neq 0$, the distribution of r is very skewed. When ρ is greater than zero, the skewness tends toward the left, with high values of r being relatively more probable than lower values. The skewness is reversed for ρ less than zero. This complicated dependency of the sampling distribution of r on the value of ρ makes it impossible to employ the t-test or normal distribution directly. To permit inferences about $\rho \neq 0$, R. A. Fisher developed for a bivariate normal population the Z-transformation given by

$$Z = 0.5 \ln \left[\frac{1 + r}{1 - r} \right] \tag{139}$$

For sample correlation coefficients r computed from independent draws from a bivariate normal population whose correlation is ρ, the statistic Z is approximately normally distributed, with a mean given by

$$\mu_Z = 0.5 \ln [\frac{1 + \rho}{1 - \rho}] \tag{140}$$

and a standard deviati›n given by

$$\sigma_Z = \sqrt{1/(N-3)} \tag{141}$$

where N is the sample size. The goodness of these approximations increases with smaller absolute values of ρ and with larger sample sizes N.

If the population correlation is ρ and one samples from it repeatedly, 95 percent of the sample correlation coefficients drawn from the population will fall between the so called "95-percent confidence limits" of ρ. Thus, from a single value of r that happens to lie within those limits, one can infer with only a 5-percent risk of error that the population correlation is ρ. More precisely, it can be said with only a 5-percent risk that r is not significantly different from the stated ρ.

ENDs generated by the V2S1 model have the bivariate normal distribution, but if all the simulated ceiling and visibility observations produced by V2S1 are included in the sample used for calculating correlations, then the data have not been sampled independently, and a correction must be made accordingly. If a correction is made for serial dependency, it is possible to make hypotheses about the correlations ρ in the V2S1 model.

Hypothesize that a population correlation of the V2S1 model is $\rho \neq 0$ and calculate the 95-percent confidence limits about ρ based on sampling the V2S1 process N consecutive times. To correct for serial dependency in the time series of V2S1 observations, Equation (60) of the AWS Guide for Applied Climatology (see references) is used:

$$N' = N [\frac{1 - \rho}{1 + \rho}] \tag{142}$$

where N' is the effective number of independent observations in a sample of size N. Then Fisher's Z-statistic is calculated using Equation (140) and the standard deviation from

$$\sigma_z = \sqrt{1/(N'-3)} \tag{143}$$

The 95-percent confidence limits in Z are given by

$$Z_u = Z + 1.96\sigma_z \tag{144}$$

$$Z_l = Z - 1.96\sigma_z \tag{145}$$

An inverse Fisher Z-transformation is used to convert the confidence limits in the Z domain to confidence limits in the ρ domain

$$\rho_u = \frac{\exp(\ 2Z_u\ ) - 1}{\exp(\ 2Z_u\ ) + 1} \tag{146}$$

$$\rho_l = \frac{\exp(\ 2Z_l\ ) - 1}{\exp(\ 2Z_l\ ) + 1} \tag{147}$$

Fisher's Z-transform can also be expressed in the form of the hyperbolic tangent. Multiplying the numerator and denominator of Equation (146) by e^{-Z} gives

$$\rho = \frac{e^Z - e^{-Z}}{e^Z + e^{-Z}} \approx \tanh\ (Z) \tag{148}$$

$$Z = \tanh^{-1}\ (\rho) \tag{149}$$

Results in Table 15 show that the sample correlation coefficients produced by the V2S1 model fall within the 95-percent confidence limits of the hypothesized correlation. Hence, the model appears to preserve the serial correlation ρ_{cc} of the ceiling, the serial correlation ρ_{vv} of the visibility, and the cross-correlation ρ_{cv} of ceiling and visibility. In addition, the cross-lag correlation $\rho_{v_t c_{t+1}}$ does appear to reduce to the automatic correlation $\rho_{cc}\rho_{cv}$.

3.4.2 Marginal Distributions. If a particular model of the cumulative distribution functions of ceiling and visibility is used in the transnormalization process of the V2S1 model, then a long run of that model should return the same distributions, within the accuracy limits imposed by the sampling error. To test this, Somerville and Bean's ceiling and visibility models for 1200 LST, February, at Scott AFB, IL, were used in transnormalization. The V2S1 model was then exercised over a long run of 100,000 observations, each falling at 1200 LST (because a 24-hour time step was used). The month was restricted to February. Results, shown in Table 16, indicate that the V2S1 modeı does preserve the marginal distributions of ceiling and visibility, within the limits of accuracy imposed by sampling error.

Table 16. Marginal Distribution Tests.

N = 100,000 Observations

	Cumulative Distribution Function			Cumulative Distribution Function	
Ceiling (ft)	Bean and Somerville	V2S1	Visibility (SM)	Somerville and Bean	V2S1
20,000	0.458	0.458	6.0	0.683	0.682
10,000	0.547	0.548	4.0	0.826	0.829
3,000	0.733	0.733	3.0	0.890	0.892
2,000	0.798	0.799	2.0	0.944	0.944
1,000	0.895	0.897	1.0	0.982	0.983
200	0.989	0.989	0.5	0.995	0.995
0	1.000	1.000	0	1.000	1.000

3.4.3 Synthetic RUSSWOs. By adjusting the cross-correlation ρ_{CV}, it is possible to adjust the joint probabilities of ceiling and visibility produced by the V2S1 model and thus to produce -- either analytically or by Monte Carlo simulation -- synthetic RUSSWOs tuned to match actual RUSSWOs. If ENDs of actual weather variables were distributed exactly according to a multivariate normal distribution, and if no bias were introduced by the method used to observe and record the weather, then the V2S1 model could produce synthetic RUSSWOs differing from "natural" RUSSWOs by no more than sampling error.

In practice, weather observations contain biases and inaccuracies that are at least as bad as assuming the ENDs of these data are multivariate normal. Thus, three sources of error -- observing/recording bias, non-multinormality and sampling error -- complicate the process of "tuning" V2S1 to reproduce a particular RUSSWO. Even if a nearly perfect fit were attained between a synthetic and a "natural" RUSSWO, one would merely be tempted to ask, "Have you fitted nature or have you fitted an inadequate perception of nature?"

Putting aside the question of the basic advisability of "fitting" a synthetic RUSSWO to a natural RUSSWO, it is remarkable how close a fit can be obtained just by "tuning" the cross-correlation ρ_{CV}. Table 17 presents an example for Scott AFB, Illinois, in February at 1200 LST. The largest differences between the synthetic and the natural RUSSWO is 0.038.

Table 17. Comparison of the Scott AFB, IL RUSSWO and Joint Probability of Ceiling and Visibility Produced by the V2S1 Model.

CEILING VS VISIBILITY SECTION, RUSSWO, SCOTT AFB, IL, FEB, 12-14L, EXTRACT

	Visibility (statute miles)						
Ceiling (ft)	6.0	4.0	3.0	2.0	1.0	0.5	0.0
20,000	0.448	0.483	0.488	0.491	0.491	0.491	0.491
10,000	0.506	0.548	0.556	0.560	0.560	0.560	0.560
3,000	0.601	0.669	0.686	0.692	0.695	0.695	0.695
2,000	0.656	0.749	0.771	0.780	0.780	0.787	0.787
1,000	0.689	0.817	0.857	0.876	0.892	0.896	0.896
200	0.695	0.836	0.888	0.928	0.980	0.998	1.000
0	0.695	0.836	0.888	0.928	0.980	0.998	1.000

Synthetic RUSSWO Produced by V2S1 Model, Scott AFB, IL, 12L

Time Step Δt = 24 hr ρ_{cv} = 0.72 No Recording Mask Total Obs = 100,000

	Visibility (statute miles)						
Ceiling (ft)	6.0	4.0	3.0	2.0	1.0	0.5	0.0
20,000	0.420	0.448	0.454	0.457	0.458	0.458	0.458
10,000	0.488	0.530	0.541	0.546	0.548	0.548	0.548
3,000	0.603	0.685	0.710	0.725	0.732	0.733	0.733
2,000	0.634	0.734	0.766	0.787	0.797	0.799	0.799
1,000	0.667	0.793	0.841	0.874	0.893	0.896	0.897
200	0.682	0.828	0.890	0.940	0.976	0.986	0.989
0	0.682	0.829	0.892	0.944	0.983	0.995	1.000

3.5 Summary and Conclusions

A single-station, two-variable model has been constructed and tested. The model has been found to preserve the serial correlation ρ_{cc} of ceiling over time, the serial correlation ρ_{vv} of visibility and the cross-correlation ρ_{cv} of ceiling and visibility. The cross-lag correlation $\rho_{v_t c_{t+1}}$ in this model reduces to the automatic correlation $\rho_{cc}\rho_{vv}$. Furthermore, the model appears to preserve the marginal distributions of ceiling and visibility as well as to give a faithful representation of the joint probabilities between ceiling and visibility, as shown in the interior of a RUSSWO.

Models of the cumulative distribution functions of ceiling and visibility, developed by Somerville and Bean, have been found superior to others tested. USAFETAC currently has the capability for fitting the Weibull curve to visibility data from anywhere on the globe and has a limited capability to fit Burr curves.

Chapter 4

MULTIVARIABLE/MULTISTATION MODELS

4.1 General

Although the V1S1 and V2S1 models have been shown to be excellent for time series of one or two variables, few simulation support requests are simple enough to be served by these models. In general, users of environmental simulation models want simulation techniques that are not limited to two variables. In this chapter, a multivariate triangular matrix model capable of generating a large number of correlated elements will be discussed. These elements could represent several variables at a single station or a single variable at multiple locations. Thus, this type of environmental simulation model allows more flexibility than the V1S1 and V2S1 models.

4.2 Generation of Random Normal Vectors with Desired Correlation Using the Multivariate Triangular Matrix Model (MULTRI)

Let $\underline{X}$ be a vector stochastic variable consisting of $j = 1, 2, 3, \ldots, M$ scalar variables X_j and $k = 1, 2, 3, \ldots, N$ observations $\underline{X}_k$ thus,

$$\underline{X} = [X_{ik}] = \begin{array}{c|cccc|} & & j\rightarrow & & \\ & X_{11} & X_{12} & \cdots & X_{1M} \\ k & X_{21} & X_{22} & \cdots & X_{2M} \\ \downarrow & \cdots & \cdots & \cdots & \cdots \\ & X_{N1} & X_{N2} & \cdots & X_{NM} \end{array} \qquad (150)$$

The kth observation of $\underline{X}$ is thus the row vector,

$$\underline{X}_k = [X_{1k}\ X_{2k}\ \cdots\ X_{Mk}] \qquad (151)$$

The vector of means is

$$E(\underline{X}_k) = [\bar{X}_1 \quad \bar{X}_2 \quad \ldots \quad \bar{X}_M] \qquad (152)$$

which may also be shown as an (N x M) matrix all of whose rows are identical.

The random variable $\underline{X}$ can be expressed in terms of its deviation from the mean $\bar{\underline{X}}$ by

$$\underline{x} = \underline{X} - \bar{\underline{X}} \qquad (153)$$

The sum of the squares and cross products (SSCP) in raw-score form is the symmetric matrix $\underline{X}'\underline{X}$. Since any particular observation $\underline{X}_k$ is a (1 x M) row vector, the raw-score SSCP is a

$$(M \times 1) \times (1 \times M) = (M \times M) - \text{dimensional}$$

matrix given by

$$\underline{X}'\underline{X} = \begin{vmatrix} \Sigma X_1{}^2 & \Sigma X_1 X_2 & \cdots & \Sigma X_1 X_M \\ \Sigma X_2 X_1 & \Sigma X_2{}^2 & \cdots & \Sigma X_2 X_M \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma X_M X_1 & \Sigma X_M X_2 & \cdots & \Sigma X_M{}^2 \end{vmatrix} \qquad (154)$$

Similarly the deviation-score SSCP is

$$\underline{x}'\underline{x} = \begin{vmatrix} \Sigma x_1{}^2 & \Sigma x_1 x_2 & \cdots & \Sigma x_1 x_M \\ \Sigma x_2 x_1 & \Sigma x_2{}^2 & \cdots & \Sigma x_2 x_M \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma x_M x_1 & \Sigma x_M x_2 & \cdots & \Sigma x_M{}^2 \end{vmatrix} \qquad (155)$$

The two are related by (Tatsuoka, 1971)

$$\underline{x}'\underline{x} = \underline{X}'\underline{X} - \underline{\bar{X}}'\underline{\bar{X}} \qquad (156)$$

which gives the computational rule for obtaining $\underline{x}'\underline{x}$.

An unbiased estimate of the dispersion or variance-covariance matrix $\underline{D}$ is given by dividing the elements of the deviation-score SSCP by the number of degrees of freedom, i.e., N - 1.

$$\underline{D} = \frac{1}{N-1} \sum_{k=1}^{N} \underline{x}_k'\underline{x}_k \qquad (157)$$

or

$$D_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} x_{ik} x_{jk} \qquad (157a)$$

where k is a datum index varying from k = 1 for the first vector x to k = N for the final vector. Note that $\underline{D}$ is a symmetric (M x M)-dimensional matrix.

The maximum likelihood estimate of $\underline{D}$ is given by

$$\underline{D} = E(\underline{x}'\underline{x}) = \frac{1}{N}\sum_{k=1}^{N} \underline{x}_k'\underline{x}_k \tag{158}$$

or

$$D_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ik}x_{jk} \tag{158a}$$

The maximum likelihood estimate is used in this and similar contexts because as long as N' data are independent out of N total data, the variance-covariance matrix will be positive definite. Such a matrix is, in theory, invertible. It should be kept in mind that the maximum likelihood estimate is biased; variance-covariance estimates will be smaller, on the average, than they should oe. The bias is not a problem in this application.

The variance of a variable X is

$$\begin{aligned}\sigma_X{}^2 &= E[\ (X-\mu_X)^2\ ] \\ &= E[\ (X-\mu_X)(X-\mu_X)\ ] \\ &= \frac{1}{N}\sum_{k=1}^{N} (X_k-\mu_X)(X_k-\mu_X)\end{aligned} \tag{159}$$

The covariance between two variables X and Y is

$$\begin{aligned}\sigma_{XY} &= E[\ (X-\mu_X)(Y-\mu_Y)\ ] \\ &= \frac{1}{N}\sum_{k=1}^{N} (X_k-\mu_X)(Y_k-\mu_Y)\end{aligned} \tag{160}$$

Recalling Equation (57), the linear correlation ρ_{XY} between X and Y is simply the covariance between X and Y divided by the product of the standard deviations of X and Y, i.e.,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\ \sigma_Y}$$

or

$$\rho_{XY} = E[\ (\frac{X-\mu_X}{\sigma_X})\ (\frac{Y-\mu_Y}{\sigma_Y})\ ] \tag{161}$$

Note that the covariance of X with X reduces to the variance of X, i.e.,

$$\sigma_{XX} = E[\ (X-\mu_X)(X-\mu_X)\ ] = E[\ (X-\mu_X)^2\ ] = \sigma_X{}^2 \tag{162}$$

The covariance between a variable X_1 and a variable X_2 is

$$\sigma_{12} = \frac{1}{N}\sum_{k=1}^{N}(X_1-\bar{X}_1)(X_2-\bar{X}_2) = \frac{1}{N}\sum_{k=1}^{N}X_1X_2 \tag{163}$$

and between X_1 and X_3 is

$$\sigma_{13} = \frac{1}{N}\sum_{k=1}^{N}X_1X_3 \tag{164}$$

Thus, the variance-covariance matrix $\underline{D}$ can be written as

$$\underline{D} = \begin{vmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{vmatrix} \tag{165}$$

or

$$\underline{D} = \begin{vmatrix} \sigma_1{}^2 & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_2{}^2 & \cdots & \sigma_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_M{}^2 \end{vmatrix} \tag{156}$$

Because

$$\sigma_{XY} = \sigma_X\ \sigma_Y\ \rho_{XY} \tag{167}$$

the variance-covariance matrix can be written as

$$\underline{D} = \begin{vmatrix} \sigma_1{}^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_M\rho_{1M} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2{}^2 & \cdots & \sigma_2\sigma_M\rho_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_M\sigma_1\rho_{M1} & \sigma_M\sigma_2\rho_{M2} & \cdots & \sigma_M{}^2 \end{vmatrix} \tag{168}$$

The form of the variance-covariance matrix $\underline{D}$ is such that the sample variances $\sigma_i{}^2$ are along the main diagonal and the sample covariances σ_{ij}, $i \neq j$ are the off-diagonal elements.

In the special case of a random variable $\underline{X}$ distributed normally with a mean of zero and a variance of one, i.e., N($\underline{0}$,1),

$$\bar{\underline{X}} = \underline{0} \tag{169}$$

$$\underline{x} = \underline{X} \tag{170}$$

$$\underline{x}'\underline{x} = \underline{X}'\underline{X} \tag{171}$$

and

$$\underline{D} = \underline{R} \tag{172}$$

where $\underline{R}$ is the correlation matrix, given by

$$\underline{R} = \begin{vmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{21} & 1 & \cdots & \rho_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{M1} & \rho_{M2} & \cdots & 1 \end{vmatrix} \tag{173}$$

which is symmetric.

If the vector stochastic variable $\underline{X}$ has the multivariate normal probability distribution, then the probability density function of $\underline{X}$ is

$$f(\underline{X}) = \frac{1}{\sqrt{|2\pi\underline{D}|}} \exp[\ -\tfrac{1}{2}(\underline{X}-\mu_X)'\underline{D}^{-1}(\underline{X}-\mu_X)\ ] \tag{174}$$

where | | represents the determinant and $\underline{D}^{-1}$ is the inverse of the variance-covariance matrix $\underline{D}$.

Random multivariate normal vectors $\underline{X}$ with a mean vector μ_X and variance-covariance matrix $\underline{D}$ can be generated by using a theorem (Anderson, 1958) which states that if η is a standard normal vector containing independent normal variable components η_i each distributed N(0,1), then there exists a unique lower triangular matrix $\underline{C}$ such that

$$\underline{X} = \underline{C}\ \eta + \mu_X \tag{175}$$

where $\underline{C}$ is an (M x M) matrix and $\underline{X}$ and μ_X are (M x 1) column vectors. Here, $\eta = [\eta_i]$ can be formed by selecting random normal numbers from a population distributed N(0,1). In this case ($\underline{X}$ - μ_X) has the (M x M) variance-covariance matrix,

$$\underline{D} = \underline{C}'\underline{C} \tag{176}$$

and the generation matrix $\underline{C}$ is obtained by a lower triangularization of the desired variance-covariance matrix $\underline{D}$. The important point is that the components of the vector $\underline{X}$ generated by this algorithm can have any desired correlation, as provided in the variance-covariance matrix $\underline{D}$, and can have any mean, as provided in the vector μ_X. By this method it is possible to generate correlated random normal numbers. If the covariances of $\underline{D}$ are zero, the elements of the generated $\underline{X}$ are then uncorrelated, i.e., independent.

One way of triangularizing the (M x M) variance-covariance matrix $\underline{D}$ to obtain the (M x M) lower triangular matrix $\underline{C}$ is the Cholesky or so called "square-root method" described in Section 4.3 of this report. Consider a case in which it is desired to generate $\underline{X}$ in three components,

$$\underline{X} = [X_1 \quad X_2 \quad X_3] \tag{177}$$

with mean

$$\underline{\mu}_X = [\mu_1 \quad \mu_2 \quad \mu_3] \tag{178}$$

and variances and covariances given by

$$\underline{D} = \begin{vmatrix} \sigma_1{}^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2{}^2 & \sigma_2\sigma_3\rho_{23} \\ \sigma_3\sigma_1\rho_{31} & \sigma_3\sigma_2\rho_{32} & \sigma_3{}^2 \end{vmatrix} \tag{179}$$

In this case,

$$\underline{C} = \begin{vmatrix} \sigma_1 & 0 & 0 \\ \sigma_2\rho_{21} & \sigma_2\sqrt{1-\rho_{21}{}^2} & 0 \\ \sigma_3\rho_{31} & \sigma_2\,\dfrac{\rho_{32}-\rho_{31}\rho_{21}}{\sqrt{1-\rho_{21}{}^2}} & \sigma_3\sqrt{1-\rho_{31}{}^2-\dfrac{(\rho_{32}-\rho_{31}\rho_{21})^2}{(1-\rho_{21})^2}} \end{vmatrix} \tag{180}$$

The generation algorithm (Equation 175) is

$$\underline{X} = \begin{vmatrix} X_1 \\ X_2 \\ X_3 \end{vmatrix}' = \begin{vmatrix} \sigma_1\eta_1 \\ \sigma_2\rho_{21}\eta_1 + \sigma_2\sqrt{1-\rho_{21}{}^2}\,\eta_2 \\ \sigma_3\rho_{31}\eta_1 + \sigma_3\dfrac{\rho_{32}-\rho_{31}\rho_{21}}{\sqrt{1-\rho_{21}{}^2}}\eta_2 + \sigma_3\sqrt{1-\rho_{31}{}^2-\dfrac{(\rho_{32}-\rho_{31}\rho_{21})^2}{(1-\rho_{21}{}^2)}}\,\eta_3 \end{vmatrix}' + \begin{vmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{vmatrix}' \tag{181}$$

In the special case where it is desired to generate $\underline{X}$ with a mean of zero and a variance of one, i.e., distributed N(0,1),

$$\underline{\mu} = \underline{0} \qquad\qquad \underline{\sigma} = \underline{1} \tag{182}$$

and the generation algorithm reduces to

$$\underline{X} = \begin{vmatrix} X_1 \\ X_2 \\ X_3 \end{vmatrix}' = \begin{vmatrix} \eta_1 \\ \rho_{21}\eta_1 + \sqrt{1-\rho_{21}^2}\,\eta_2 \\ \rho_{31}\eta_1 + \dfrac{\rho_{32}-\rho_{31}\rho_{21}}{\sqrt{1-\rho_{21}^2}}\,\eta_2 + \sqrt{1 - \rho_{31}^2 - \dfrac{(\rho_{32}-\rho_{31}\rho_{21})^2}{(1-\rho_{21}^2)}}\,\eta_3 \end{vmatrix}' \tag{183}$$

where η_1, η_2, and η_3 are numbers drawn independently from a population distributed normally with a mean of zero and variance of one.

In practice, these analytic expressions for the lower triangular matrix $\underline{C}$ are not needed. One simply forms the desired variance-covariance matrix $\underline{D}$ (or the correlation matrix $\underline{R}$ if $\underline{X}$ is to be distributed N(0,1)), lower triangularizes that matrix by the Cholesky procedure (see Section 4.3), and uses it and the mean vector $\underline{\mu}_X$ in the generation algorithm (Equation 175). The independent random normal numbers $\underline{\eta}$ are produced either by using a pseudo-random normal number generator directly or by using a uniform pseudo-random number generator and any of several suitable transformations (Naylor, et al., 1966).

The generation algorithm of Equation (175) can be illustrated with a test case. Suppose it is desired to generate a vector,

$$\underline{X} = [X_1\ X_2\ X_3\ X_4]$$

of standard normal variables (distributed N(0,1)) having the correlation matrix,

$$\underline{R} = \underline{D} = \begin{vmatrix} 1.0 & 0.8 & 0.7 & 0.3 \\ 0.8 & 1.0 & 0.6 & 0.4 \\ 0.7 & 0.6 & 1.0 & 0.5 \\ 0.3 & 0.4 & 0.5 & 1.0 \end{vmatrix}$$

The Cholesky reduction procedure is used to find the lower triangular matrix $\underline{C}$,

$$\underline{C} = \begin{vmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.8000 & 0.6000 & 0.0000 & 0.0000 \\ 0.7000 & 0.0667 & 0.7110 & 0.0000 \\ 0.3000 & 0.2667 & 0.3829 & 0.8321 \end{vmatrix}$$

The transpose of $\underline{C}$ is

$$\underline{C}' = \begin{vmatrix} 1.0000 & 0.8000 & 0.7000 & 0.3000 \\ 0.0000 & 0.6000 & 0.0667 & 0.2667 \\ 0.0000 & 0.0000 & 0.7110 & 0.3829 \\ 0.0000 & 0.0000 & 0.0000 & 0.8321 \end{vmatrix}$$

from which it can be verified that

$$\underline{C}'\underline{C} = \underline{D}$$

The matrix $\underline{C}$ is then used to generate successive values of $\underline{X}$ by performing the matrix-vector multiplication of Equation (175) with successive values of $\underline{\eta}$.

4.3 Cholesky or "Square Root" Factorization

A square matrix $\underline{A}$ whose leading submatrices have nonzero determinants can be factored (non-uniquely) as

$$\underline{A} = \underline{L}_1 \, \underline{U}_1 \qquad \text{(LU Theorem)} \qquad (184)$$

where $\underline{L}_1$ and $\underline{U}_1$ are lower and upper triangular, respectively. Likewise, $\underline{A}$ may be factored (uniquely) as

$$\underline{A} = \underline{L} \, \underline{U}_2 \qquad (185)$$

where $\underline{L}$ is a lower triangular matrix whose diagonal elements are all unity, and $\underline{U}_2$ is an upper triangular matrix. The matrix $\underline{U}_2$ can also be factored as

$$\underline{U}_2 = \underline{D} \, \underline{U} \qquad (186)$$

where $\underline{U}$ is an upper triangular matrix whose diagonal elements are all unity, and $\underline{D}$ is a diagonal matrix whose elements are the corresponding elements of $\underline{U}_2$, i.e.,

$$\underline{D} = \text{diag}\ [D_1, \ D_2, \ \ldots \ D_N] \qquad (187)$$

Using Equation (186) in Equation (185),

$$\underline{A} = \underline{L} \, \underline{D} \, \underline{U} \qquad \text{(LDU Theorem)} \qquad (188)$$

If $\underline{A}$ is symmetric and positive definite (a matrix $\underline{A}$ of order n is positive definite if $\underline{x}' \, \underline{A} \, \underline{x} > 0$, for every real, nonzero n-vector $\underline{x}$),

$$\underline{A} = \underline{L} \, \underline{D} \, \underline{L}' \qquad (189)$$

where $\underline{L}'$ is the transpose of L. Hence, $\underline{U} = \underline{L}'$ and $\underline{A}$ can be factored as

$$\underline{A} = (\underline{L}\ \underline{D}^{1/2})\ (\underline{D}^{1/2}\ \underline{L}') \qquad (190)$$

where

$$\underline{D}^{1/2} = \mathrm{diag}\ [D_1^{1/2}\ \ D_2^{1/2}\ \ \ldots\ \ D_N^{1/2}] \qquad (191)$$

It is convenient to define

$$S = L\ D^{1/2} \qquad (192)$$

so that

$$\underline{A} = \underline{S}\ \underline{S}' \qquad (193)$$

Therefore, any real, symmetric, positive definite matrix $\underline{A}$ can be factored into a lower triangular matrix $\underline{S}$ and its transpose $\underline{S}'$.

The algorithm of choice to perform the factorization of Equation (193) is the Cholesky or so called square-root method (Acton, 1970; Carnahan, et al., 1969; Forsythe, et al., 1967; Naylor, et al., 1966; and Scheuer and Stoller, 1962). The Cholesky method is extremely stable, never requires interchanging to avoid small pivots, and requires the least computational labor of all decomposition schemes, largely because of the symmetry of the $\underline{A}$ matrix. If the symmetric, positive definite requirements are not adhered to, the Cholesky or square-root algorithm will break down by calling for division by zero or attempting to take the square root of a negative number.

The Cholesky or square-root algorithm for factoring the real, symmetric, positive definite matrix $\underline{A} = [a_{ij}]$ of order n into a lower triangular matrix $\underline{S} = [s_{ij}]$ and its transpose consists of three rules

$$s_{i1} = \frac{a_{i1}}{\sqrt{a_{11}}} \qquad \begin{matrix} j = 1 \\ 1 \leq i \leq n \end{matrix} \qquad (194)$$

$$s_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} s_{ik}^2} \qquad \begin{matrix} j > 1 \\ 1 < i \leq n \end{matrix} \qquad (195)$$

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} s_{ik}s_{jk}}{s_{jj}} \qquad \begin{matrix} j > 1 \\ 1 < j < i \leq n \end{matrix} \qquad (196)$$

Finally, $s_{ij} = 0$ for all $j > i$. These rules are implemented column-wise, starting with the leftmost column ($j = 1$) and proceeding down each column (toward increasing i). This becomes apparent when the algorithm is written out

```
procedure LUSQRT (n, A, S)
integer i, j, k, n
real array A [1:n, 1:n], S [1:n, 1:n]
begin

    j: = 1;
    for i: = 1 step 1 until n do
    begin

        $S_{i,j}: = A_{i,j} / \text{sqrt} (A_{1,1});$

    end;
    for j: = 2 step 1 until n do
    begin

        $S_{j,j}: = \text{sqrt} (A_{j,j} - \sum_{k=1}^{j-1} S_{j,k}^2);$

        for i: = j + 1 step 1 until n do
        begin

            $S_{i,j}: = (A_{i,j} - \sum_{k=1}^{j-1} [S_{i,k} S_{j,k}]) / S_{j,j};$

        end;

    end;

end LUSQRT;
```

4.4 Derivation of Single-station, Two-variable Model Equations from the Multivariate Triangular Matrix Model (MULTRI)

The V2S1 single-station, two-variable model described in Chapter 3 is actually a special case of the multivariate triangular matrix model (MULTRI), namely, the case where the number of variables is two. It should therefore be possible to derive the equations of the single-station, two-variable model, (V2S1), q.v.,

$$\ddot{c}_t = \rho_{cc}\ddot{c}_0 + \sqrt{1-\rho_{cc}^2}\,\eta_c \tag{120}$$

$$\ddot{v}_t = \rho_{vv}\ddot{v}_0 + \sqrt{1-\rho_{vv}^2}\,\rho'_{cv}\eta_c + \sqrt{1-\rho_{vv}^2}\,\sqrt{1-\rho'^2_{cv}}\,\eta \tag{197}$$

from the triangular matrix formulation. Recalling Equation (129),

$$\rho'_{cv} = \frac{1-\rho_{cc}\rho_{vv}}{\sqrt{1-\rho_{cc}^2}\,\sqrt{1-\rho_{vv}^2}}\,\rho_{cv} \tag{129}$$

Using Equation (129) in Equation (197) produces the altered form,

$$\ddot{v}_t = \rho_{vv}\ddot{v}_0 + \frac{(1-\rho_{cc}\rho_{vv})}{\sqrt{1-\rho_{cc}^2}}\,\rho_{cv}\eta_c + \sqrt{(1-\rho_{vv}^2) - \frac{(1-\rho_{cc}\rho_{vv}^2)\rho_{cv}^2}{(1-\rho_{cc}^2)}}\;\eta \tag{197a}$$

Together, Equations (120) and (197a) constitute the set of simulation equations used in the single-station, two-variable model V2S1.

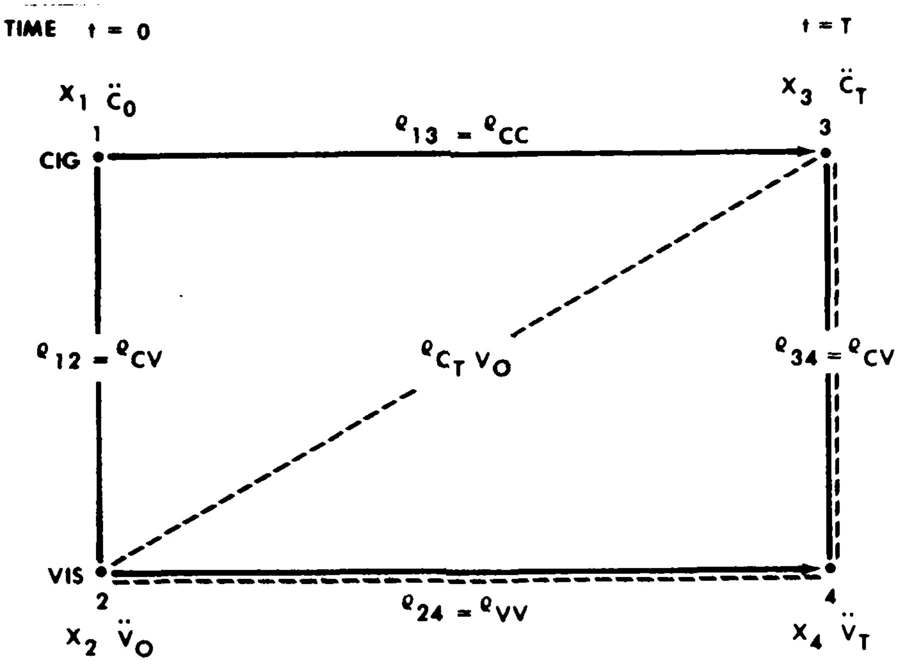


Figure 11. Correlation Influence Diagram for Single-station, Multiparameter Model.

The process used in the V2S1 model is shown in the correlation influence diagram depicted in Figure 11. In this diagram, the states of ceiling and visibility are numbered as well as lettered to show the correspondence between triangular matrix states X_1, X_2, X_3, X_4 with multiparameter model states $\ddot{c}_0$, $\ddot{v}_0$, $\ddot{c}_t$, and $\ddot{v}_t$, respectively, i.e.,

$$
\begin{aligned}
X_1 &= \ddot{c}_0 \qquad & X_3 &= \ddot{c}_t \\
X_2 &= \ddot{v}_0 \qquad & X_4 &= \ddot{v}_t
\end{aligned}
\tag{198}
$$

In order for the triangular matrix model to resemble the two-variable model V2S1, the following correlation structure is needed

$$
\begin{aligned}
\rho_{13} &= \rho_{31} = \rho_{cc} \\
\rho_{24} &= \rho_{42} = \rho_{vv} \\
\rho_{12} &= \rho_{21} = \rho_{cv} \\
\rho_{34} &= \rho_{43} = \rho_{cv}
\end{aligned}
\tag{199}
$$

The cross-lag correlations are modeled as automatic correlation

$$
\begin{aligned}
\rho_{c_t v_0} &= \rho_{cc}\ \rho_{cv} \\
\rho_{23} &= \rho_{21}\ \rho_{13}
\end{aligned}
\tag{200}
$$

$$\rho_{v_t c_0} = \rho_{vv}\,\rho_{cv}$$

$$\rho_{14} = \rho_{12}\,\rho_{24} \tag{201}$$

Under this correlation structure, it is not in general true, that ρ_{23} is equal to $\rho_{23}\rho_{43}$ or that ρ_{14} is equal to $\rho_{13}\rho_{34}$. This would be true only if $\rho_{cc} = \rho_{vv}$.

The triangular matrix model is based on a variance-covariance matrix $\underline{D}$ such as

$$\underline{D} = \left|\begin{matrix} \sigma_1{}^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} & \sigma_1\sigma_4\rho_{14} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2{}^2 & \sigma_2\sigma_3\rho_{23} & \sigma_2\sigma_4\rho_{24} \\ \sigma_3\sigma_1\rho_{31} & \sigma_3\sigma_2\rho_{32} & \sigma_3{}^2 & \sigma_3\sigma_4\rho_{34} \\ \sigma_4\sigma_1\rho_{41} & \sigma_4\sigma_2\rho_{42} & \sigma_4\sigma_3\rho_{43} & \sigma_4{}^2 \end{matrix}\right| \tag{202}$$

Since the variables generated are distributed N(0,1), the variance-covariance matrix $\underline{D}$ reduces to the correlation matrix $\underline{R}$.

$$\underline{D} = \left|\begin{matrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{matrix}\right| \tag{203}$$

$$\underline{D} = \underline{R}$$

This matrix $\underline{R}$ can be lower triangularized using the rules stated in Section 4.3 of this report dealing with Cholesky reduction. The result is a lower triangular matrix $\underline{C}$, given by

$$\underline{C} = \left|\begin{matrix} 1 & 0 & 0 & 0 \\ \rho_{21} & \sqrt{1-\rho_{21}{}^2} & 0 & 0 \\ \rho_{31} & \dfrac{\rho_{32}-\rho_{31}\rho_{21}}{\sqrt{1-\rho_{21}{}^2}} & \dfrac{\sqrt{(1-\rho_{31}{}^2)(1-\rho_{21}{}^2)-(\rho_{32}-\rho_{31}\rho_{21})^2}}{\sqrt{1-\rho_{21}{}^2}} & 0 \\ \rho_{41} & \dfrac{\rho_{42}-\rho_{41}\rho_{21}}{\sqrt{1-\rho_{21}{}^2}} & c_{43} & c_{44} \end{matrix}\right| \tag{204}$$

$$c_{43} = \frac{(\rho_{43}-\rho_{41}\rho_{31})\sqrt{1-\rho_{21}{}^2} - (\rho_{42}-\rho_{41}\rho_{21})(\rho_{32}-\rho_{31}\rho_{21})}{\sqrt{(1-\rho_{31}{}^2)(1-\rho_{21}{}^2) - (\rho_{32}-\rho_{31}\rho_{21})^2}} \tag{205}$$

$$c_{44} = \sqrt{1 - \rho_{41}{}^2 - \frac{(\rho_{42}-\rho_{41}\rho_{21})^2}{(1-\rho_{21}{}^2)} - \frac{(\rho_{43}-\rho_{41}\rho_{31})^2}{(1-\rho_{31}{}^2)}} \qquad (206)$$

Applying automatic correlation,

$$\rho_{32} - \rho_{31}\rho_{21} = 0$$

$$\underline{C} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ \rho_{21} & \sqrt{1-\rho_{21}{}^2} & 0 & 0 \\ \rho_{31} & 0 & \sqrt{1-\rho_{31}{}^2} & 0 \\ \rho_{41} & \dfrac{\rho_{42}-\rho_{41}\rho_{21}}{\sqrt{1-\rho_{21}{}^2}} & \dfrac{(\rho_{43}-\rho_{41}\rho_{31})}{\sqrt{1-\rho_{31}{}^2}} & c_{44} \end{vmatrix} \qquad (207)$$

$$c_{44} = \sqrt{1 - \rho_{41}{}^2 - \frac{(\rho_{42}-\rho_{41}\rho_{21})^2}{(1-\rho_{21}{}^2)} - \frac{(\rho_{43}-\rho_{41}\rho_{31})^2}{(1-\rho_{31}{}^2)}} \qquad (208)$$

The generation algorithm for the triangular matrix method is

$$\underline{X} = \underline{C}\ \underline{\eta}$$

in the case where the components of $\underline{X}$ are standard normal variables. In that case,

$$\begin{aligned} \overset{''}{c}_0 &= X_1 = c_{1,1}\ \eta_1 \\ \overset{''}{v}_0 &= X_2 = c_{2,1}\ \eta_1 + c_{2,2}\ \eta_2 \\ \overset{''}{c}_t &= X_3 = c_{3,1}\ \eta_1 + c_{3,2}\ \eta_2 + c_{3,3}\ \eta_3 \\ \overset{''}{v}_t &= X_4 = c_{4,1}\ \eta_1 + c_{4,2}\ \eta_2 + c_{4,3}\ \eta_3 + c_{4,4}\ \eta_4 \end{aligned} \qquad (209)$$

The first two equations are used to define the lag variables $\overset{''}{c}_0$ and $\overset{''}{v}_0$ and thus to add time stepping to the model. These equations are

$$\overset{''}{c}_0 = X_1 = \eta_1 \qquad (210)$$

$$\begin{aligned} \overset{''}{v}_0 = X_2 &= \rho_{21}\eta_1 + \sqrt{1-\rho_{21}{}^2}\ \eta_2 \\ &= \rho_{cv}\overset{''}{c}_0 + \sqrt{1-\rho_{cv}{}^2}\ \eta_2 \end{aligned} \qquad (211)$$

The next two equations from (209) define the process

$$\begin{aligned}\overset{''}{c}_t = X_3 &= \rho_{31}X_1 + \sqrt{1-\rho_{31}^2}\;\eta_3 \\ &= \rho_{cc}\overset{''}{c}_0 + \sqrt{1-\rho_{cc}^2}\;\eta_c \end{aligned} \tag{212}$$

$$\begin{aligned}\overset{''}{v}_t = X_4 &= \rho_{41}\,X_1 + \frac{\rho_{42}-\rho_{41}\rho_{21}}{\sqrt{1-\rho_{21}^2}}\,\eta_2 + \frac{\rho_{43}-\rho_{41}\rho_{31}}{\sqrt{1-\rho_{31}^2}}\,\eta_3 \\ &+ \sqrt{1-\rho_{41}^2 - \frac{(\rho_{42}-\rho_{41}\rho_{21})^2}{(1-\rho_{21}^2)} - \frac{(\rho_{43}-\rho_{41}\rho_{31})^2}{(1-\rho_{31}^2)}}\;\eta_4 \end{aligned} \tag{213}$$

Equation (213) is not in final form. The variables X_1 and η_2 must be replaced by forms containing $\overset{''}{c}_0$ and $\overset{''}{v}_0$. Equation 198 can be used for X_1. Equation (211) solved for η_2 can be used to materialize $\overset{''}{v}_0$ in Equation (213)

$$\eta_2 = \frac{\overset{''}{v}_0 - \rho_{cv}\overset{''}{c}_0}{\sqrt{1-\rho_{cv}^2}} \tag{211a}$$

After these substitutions, Equation (213) becomes

$$\begin{aligned}\overset{''}{v}_t = X_4 &= \rho_{vv}\overset{''}{v}_0 + \frac{(1-\rho_{vv}\rho_{cc})}{\sqrt{1-\rho_{cc}^2}}\,\rho_{cv}\,\eta_c \\ &+ \sqrt{1-\rho_{vv}^2\rho_{cc}^2 - \frac{(\rho_{vv}-\rho_{vv}\rho_{cv}^2)^2}{(1-\rho_{cv}^2)} - \frac{(\rho_{cv}-\rho_{vv}\rho_{cv}\rho_{cc})^2}{(1-\rho_{cc}^2)}}\;\eta \end{aligned} \tag{214}$$

Some algebraic manipulation of Equation (214) produces

$$\begin{aligned}\overset{''}{v}_t = X_4 &= \rho_{vv}\,\overset{''}{v}_0 + \frac{(1-\rho_{vv}\rho_{cc})}{\sqrt{1-\rho_{cc}^2}}\,\rho_{cv}\eta_c \\ &+ \sqrt{(1-r_{vv}^2) - \frac{(1-\rho_{cc}\rho_{vv})^2\rho_{cv}^2}{(1-\rho_{cc}^2)}}\;\eta \end{aligned} \tag{214a}$$

This equation and Equation (212) agree with Equations (120) and (197a) of the single-station, two-variable model V2S1. The two models are therefore equivalent statements of the same stochastic process.

4.5 Summary and Conclusions

In this chapter a multivariate triangular matrix method for generating an independent vector of N correlated elements has been presented. The multivariate triangular matrix model MULTRI allows simulation of more than two variables. Finally, it has been shown that the Ornstein-Uhlenbeck process for two variables and the multivariate triangular matrix technique for two variables are equivalent statements of the same stochastic process. In Chapter 5 a case study of a multi-parameter multi-location model will be presented.

Chapter 5

MODELING JOINT SKY COVER DISTRIBUTIONS A CASE STUDY USING THE MULTIVARIATE TRIANGULAR MATRIX MODEL

5.1 General

USAFETAC Project 2357 required producing joint probability tables for eight selected locations in the Soviet Union (see Table 18). The requested joint probability tables were for

- Sky cover at station pairs at a fixed time
- Sky cover at a single station at some initial time and N lag times

For example, what is the probability that any two stations would have 8/8 skycover at the same time,. or what is the probability that Moscow would have 8/8 skycover at both 1200 GMT and 1500 GMT?

Table 18. Stations Modeled in Case Study.

Site Name	WMO Station #	Latitude	Longitude
Chiganak, RS	359970	45.10 N	73.97 E
Moscow, RS	276120	55.75 N	37.57 E
Vladimar, RS	275320	56.13 N	40.38 E
Kingisepp, RS	260590	59.37 N	28.60 E
Kazan, RS	275950	55.47 N	49.18 E
Feddosiya, RS	339760	45.03 N	35.38 E
Vyborg, RS	228920	60.72 N	28.80 E
Voronezh, RS	341220	51.70 N	39.17 E

The probability that a given location will have a certain amount of cloud cover can be easily estimated from available climatological data. Estimating the probability that the given location will have a certain sky cover at two or more times or that two locations will have certain sky covers at the same time is more difficult and requires processing large amounts of data. The modeling approach is a convenient alternative because it reduces the need for data processing. Furthermore, the modeling approach has the advantage of being able to smooth through certain pathologies in the raw data, a subject discussed in Section 5.2 below.

The multivariate triangular matrix model MULTRI was determined to be well suited for this type of problem. The joint sky cover probability tables could be easily produced by generating long series of independent random vectors and tabulating the results. Each vector would contain M correlated elements distributed N(0,1). Thus, the vector elements could represent values for the ENDs of the marginal distributions of sky cover for the various locations and times if a good normalizing function for sky cover could be found. The tables could then be formed by converting the ENDs to sky cover categories using the normalizing function and then tabulating and storing the raw counts for later probability calculations. The mathematics of the triangular matrix method were discussed in detail in Chapter 4, so only the application to the joint sky cover probability modeling will be discussed in this chapter.

Some important assumptions had to be made in using this approach: (1) the marginal distributions of sky cover for the individual stations could be adequately described by some normalizing function; (2) the spatial, and temporal correlation functions for the geographic location that was to be modeled could be adequately described by some correlation model; and (3) the joint occurrences of sky cover for the various station pairs and lag times were distributed multivariate normally.

The model's final results depended on the "goodness" of these assumptions. If any one of these assumptions were bad, the final model would fail to generate joint probability tables that were representative of actual conditions. In this chapter, each assumption will be discussed as it pertained to this case study and the results will be presented and compared with actual data.

5.2 Models for Marginal Distributions of Sky Cover at Individual Stations

As the first step in generating the joint probability tables, the marginal distributions of sky cover the eight individual stations were fitted to Johnson S_B curves, described below. These unconditional probability models were needed to feed the multivariate joint probabilities model. The data used to develop these individual models were prepared by OL-A, USAFETAC, Asheville NC, for the period of record January 1973 through December 1979. Ninety-six separate distributions were fitted for each station, one for each 3-hour period of the day, beginning 00-02 Local Standard Time (LST), for each of the 12 months.

The methods of Somerville were adapted to develop the models for the marginal distributions of sky cover (Somerville, Watkins, and Daley, 1978). Somerville recommends the Johnson S_B *family* of distributions because of their ease of use and the many shapes which these curves can assume. The Johnson curves are particularly useful because the modeling coefficients may be obtained by linear curve fitting or regression techniques as opposed to other functions which may require more complex nonlinear methods.

The S_B family is given by

$$\ddot{z} = \gamma + \eta \cdot \ln[\frac{x_T}{1-x_T}] \qquad (215)$$

where γ and η are the coefficients determined from empirical data. The variable x_T is some threshold value of the sky cover X in fractional coverage, and $\ddot{z}$ is the END of the cumulative distribution that the sky cover (X) is less than or equal to x_T. Equation (215) may also be solved for sky cover, given the END of the cumulative distribution

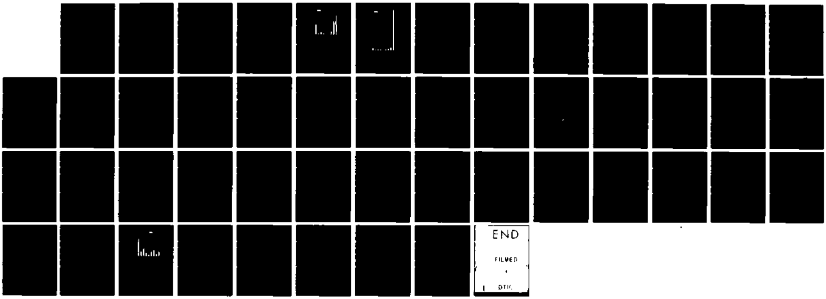
$$x_T = \frac{1 + \exp((\ddot{z}-\gamma)/\eta)}{\exp((\ddot{z}-\gamma)/\eta)} \qquad (216)$$

Using Equations (215) and (216) with particular modeling coefficients, a value for sky cover can be calculated for any value of the cumulative distribution, or a value for the cumulative distribution can be calculated for any sky cover. These equations establish a corresponding one-to-one relationship for values of sky cover and the END of the cumulative distribution for each station. This attribute lends itself quite well to stochastic modeling.

The modeling coefficients for sky cover were obtained by using a singular value decomposition (SVD) scheme to fit the observed sky cover distributions to the Johnson S_B family of curves. Singular value decomposition is described in Forsythe, Malcolm, and Moler (1977). Table 19 lists the 96 sets of coefficients for Kingisepp, RS, that were obtained during the curve fitting procedure. It is the variability of these coefficients that encompass the diurnal and seasonal variations of the sky cover distribution.

Tables 20 and 21 contain root-mean-square (RMS) difference information [illegible] each curve fit for Kingisepp, RS, and Chiganak, RS. Kingisepp represents the lowest and Chiganak the highest overall RMS values for the eight locations [illegible] were modeled. Tables 22 and 2[illegible] list the percentage of time that the m[illegible] distributions for these two stations differed from the observed at [illegible] various thresholds. The RMS information contained in these tables [illegible] validating the first assumption, that is, whether the marginal [illegible] sky cover could be adequately described by some normalizing [illegible] shows that of the 96 curve fits for Kingisepp, RS, [illegible] RMS values were greater than 3.0 and all RMS values we[illegible] be emphasized that because the observed frequency [illegible] small sample size and because of the problems [illegible] [illegible], the observed distributions that were [illegible] greater than 5 percent from the "true" [illegible] Somerville feels that in some cases [illegible] truth than the observed distributions [illegible] case, RMS values [illegible] less than 3.0 [illegible]

AD-A123 342 BASIC TECHNIQUES IN ENVIRONMENTAL SIMULATION(U) AIR 2/2
FORCE ENVIRONMENTAL TECHNICAL APPLICATIONS CENTER SCOTT
AFB IL R C WHITON ET AL. JUL 82 USAFETAC/TN-82/004
UNCLASSIFIED SBI-AD-E850 217 F/G 4/2 NL



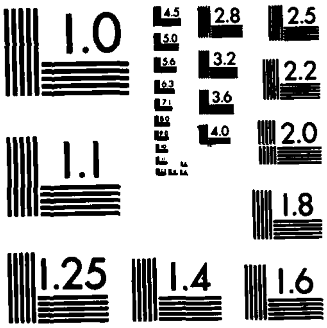

1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

Table 19. Johnson S_B Coefficients for WMO Station 260590, Kingisepp, RS.

MON/LST		0-2	3-5	6-8	9-11	12-14	15-17	18-20	21-23
JAN	γ	-0.65065733	-0.62437560	-0.70633917	-0.72740018	-0.74355437	-0.53277102	-0.48410637	-0.46169160
	η	0.07505630	0.06580009	0.05816473	0.11981313	0.18992806	0.18020875	0.15398932	0.08378725
FEB	γ	-0.31948838	-0.41016982	-0.47530818	-0.67561203	-0.62912335	-0.56448575	-0.52610560	-0.25367540
	η	0.08392397	0.05367186	0.07533571	0.15814687	0.19302432	0.22127987	0.19040769	0.08605789
MAR	γ	-0.20768025	-0.23066607	-0.46260590	-0.57092351	-0.56928215	-0.63253389	-0.52016556	-0.30937240
	η	0.06238959	0.08161962	0.07875919	0.13700320	0.22335363	0.21709767	0.20160205	0.11815375
APR	γ	0.02529001	-0.05314603	-0.38969262	-0.42305025	-0.50725450	-0.54742464	-0.31645508	-0.11925219
	η	0.14928275	0.10542282	0.17689887	0.21931967	0.28633389	0.35537389	0.31247155	0.22095223
MAY	γ	0.11070213	0.16289717	-0.06733762	-0.10521001	-0.22872997	-0.28244343	-0.13460480	-0.08231766
	η	0.19041423	0.17249830	0.21154528	0.23224430	0.28901216	0.35820847	0.31249771	0.28449846
JUN	γ	-0.02914881	-0.02667961	-0.11783740	-0.16472488	-0.32418389	-0.48175308	-0.32102821	-0.10582893
	η	0.30188084	0.23100496	0.22629208	0.27017362	0.37826689	0.41315992	0.38828502	0.36007925
JUL	γ	-0.16172350	-0.14148365	-0.34835328	-0.25696186	-0.37465594	-0.44159174	-0.42080850	-0.27677375
	η	0.27739469	0.24933507	0.24612675	0.22594158	0.36185736	0.47827210	0.43079965	0.35648259
AUG	γ	0.09085940	0.09530241	-0.35340206	-0.46652345	-0.48288327	-0.50226863	-0.32723981	-0.12929934
	η	0.20041790	0.17191060	0.23639314	0.28382791	0.39494313	0.45963899	0.40539439	0.35887634
SEP	γ	-0.13774652	-0.18704538	-0.36551304	-0.57017452	-0.56345677	-0.60902657	-0.48433585	-0.24630613
	η	0.14888814	0.14859488	0.20722733	0.25536129	0.32332978	0.36564557	0.33635892	0.23949842
OCT	γ	-0.44739673	-0.40407459	-0.39219639	-0.65295615	-0.74026336	-0.74905553	-0.72830521	-0.46520192
	η	0.16467763	0.12842901	0.12108326	0.19403304	0.23565343	0.28960738	0.26038386	0.13730433
NOV	γ	-1.06729114	-1.21891607	-1.14663155	-1.23375489	-1.37075810	-1.32047921	-1.08888307	-1.07723922
	η	0.12770539	0.14479394	0.13649477	0.17041784	0.25551441	0.30118936	0.21510619	0.13373675
DEC	γ	-0.77374867	-0.73794994	-0.92393040	-0.85400800	-0.87715279	-0.76391631	-0.75757823	-0.75970376
	η	0.11820353	0.11802477	0.08982569	0.13071656	0.17947412	0.16479992	0.12585152	0.09485495

Table 20. RMS of Individual Curve Fits for WMO Station 260590, Kingisepp, RS.

MON/LST	00	03	06	09	12	15	18	21	AVG RMS
JAN	1.19	0.77	0.68	1.06	1.61	1.65	1.46	0.77	1.15
FEB	1.75	0.75	0.76	1.74	2.22	1.46	1.73	0.96	1.42
MAR	0.92	1.44	1.17	1.33	1.81	2.13	2.80	1.32	1.61
APR	2.31	2.16	1.97	2.72	2.14	2.20	2.47	1.93	2.24
MAY	1.95	2.00	3.42	2.90	2.18	2.06	1.82	2.98	2.41
JUN	3.26	3.33	3.37	1.98	2.73	1.90	2.78	2.79	2.77
JUL	4.39	3.34	3.53	4.21	2.12	2.73	2.73	3.52	3.32
AUG	2.59	2.13	2.73	1.67	3.38	1.10	2.63	2.83	2.38
SEP	2.73	2.17	1.28	2.43	1.82	1.90	1.93	1.70	1.99
OCT	1.40	1.70	2.21	1.26	1.15	0.73	1.43	2.60	1.56
NOV	0.68	1.26	1.47	0.98	0.93	1.21	0.83	1.33	1.07
DEC	1.64	1.97	1.53	1.63	2.19	1.56	1.01	1.41	1.62

Table 21. RMS of Individual Curve Fits for WMO Station 359970, Chiganak, RS.

MON/LST	00	03	06	09	12	15	18	21	AVG RMS
JAN	2.10	1.50	1.08	1.37	3.69	4.13	3.39	1.82	2.39
FEB	1.38	1.53	1.86	4.79	4.83	2.63	3.71	1.13	2.73
MAR	1.26	1.54	1.38	2.19	3.71	3.42	3.80	1.63	2.37
APR	2.07	1.72	2.41	3.98	4.51	4.91	4.62	3.45	3.46
MAY	2.53	1.64	2.71	4.90	4.70	4.82	4.76	2.52	3.57
JUN	2.62	2.91	3.56	5.01	3.96	3.45	3.22	3.39	3.52
JUL	1.10	1.34	2.65	3.49	2.76	4.63	4.19	4.13	3.04
AUG	0.93	0.77	0.82	2.32	1.79	2.69	1.76	2.07	1.64
SEP	1.16	0.85	1.85	2.47	2.83	2.55	4.74	1.58	2.25
OCT	1.71	1.50	1.06	4.11	5.21	3.96	5.40	1.30	3.03
NOV	1.92	2.52	1.41	5.28	4.24	4.40	3.06	2.62	3.18
DEC	1.49	2.23	1.51	2.50	2.14	2.70	0.79	1.53	1.86

Table 22. Proportion of Time that the Empirical and Modeled Cumulative Distributions for Kingisepp, RS, Differ by Various Thresholds.

WMO STATION: 260590 Kingisepp, RS

MONTH	2%	5%	10%
JAN	0.047	0.000	0.000
FEB	0.203	0.000	0.000
MAR	0.234	0.016	0.000
APR	0.422	0.000	0.000
MAY	0.422	0.031	0.000
JUN	0.484	0.047	0.000
JUL	0.531	0.141	0.000
AUG	0.500	0.031	0.000
SEP	0.359	0.000	0.000
OCT	0.250	0.000	0.000
NOV	0.078	0.000	0.000
DEC	0.203	0.000	0.000
TOT	0.304	0.022	0.000

Table 23. Proportion of Time that the Empirical and Modeled Cumulative Distributions for Chiganak, RS, Differ by Various Thresholds.

WMO STATION: 359970 Chiganak, RS

MONTH	2%	5%	10%
JAN	0.375	0.063	0.000
FEB	0.391	0.094	0.000
MAR	0.391	0.063	0.000
APR	0.547	0.188	0.000
MAY	0.563	0.188	0.000
JUN	0.719	0.109	0.000
JUL	0.531	0.141	0.000
AUG	0.250	0.000	0.000
SEP	0.391	0.031	0.000
OCT	0.453	0.188	0.000
NOV	0.453	0.109	0.016
DEC	0.297	0.031	0.000
TOT	0.436	0.098	0.001

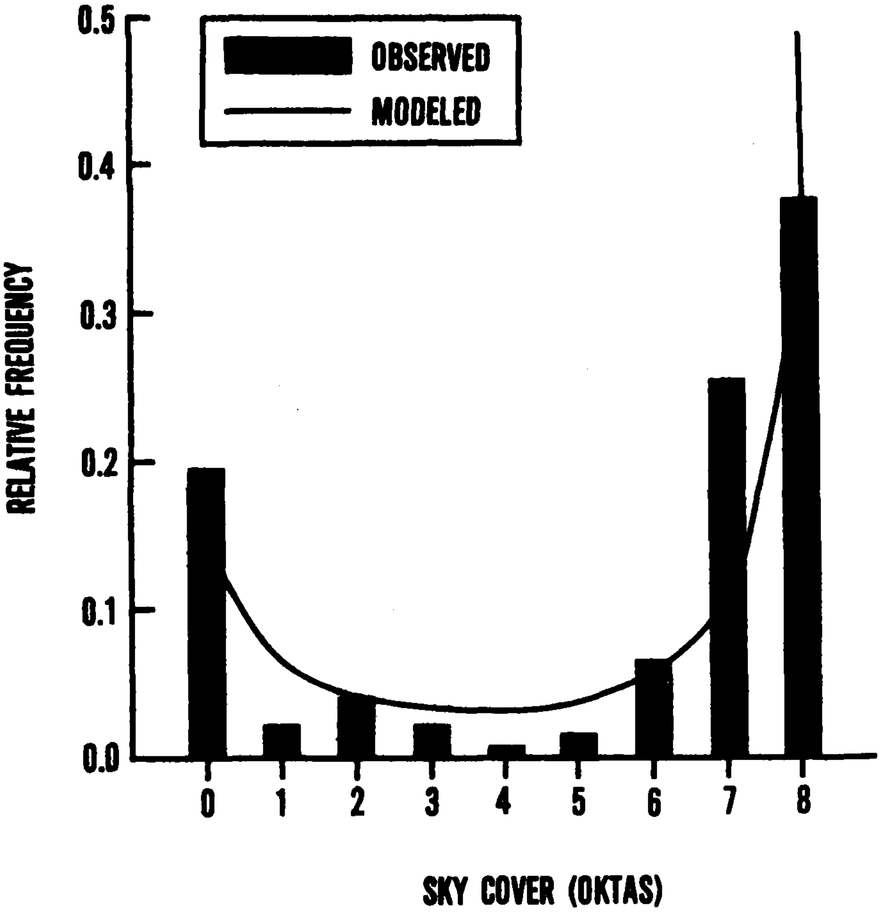


Figure 12. Relative Frequency Distribution of Cloud Cover at Chiganak, RS, November at 0900 LST. The observed distribution and the Johnson S_B curve fit to that distribution are shown. The RMS between the observed distribution and modeled CDF is 5.3 percent, and the maximum difference is 11.2 percent.

For both stations, the largest RMS values occur in the May to July period, and the lowest RMS values occur in the December to February period. A user should have more confidence in the model results from winter then those results from spring to early summer.

Figures 12 and 13 compare the modeled and observed relative frequency distributions for two individual situations. Figure 12 is for Chiganak, RS, November, 0900 LST, and represents the largest RMS value for all fits (5.3). Note that the distributions exhibit the same general shape even though the RMS is large. Figure 13 is for Moscow, RS, November, 0600 LST, and represents the smallest RMS value (0.5). In this case the modeled curve duplicated the observed distribution quite well.

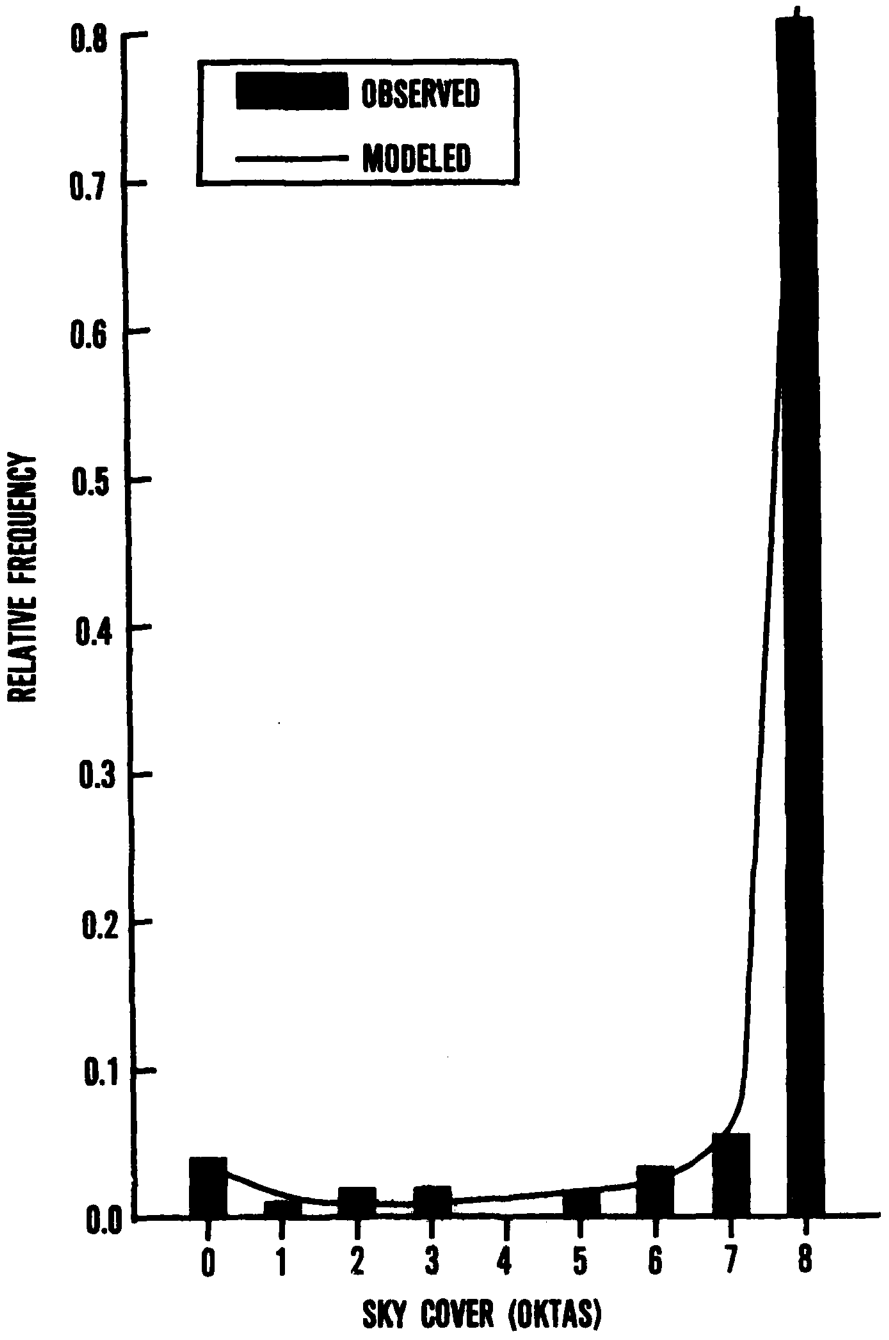


Figure 13. Relative Frequency Distribution of Cloud Cover at Moscow, RS, November at 0600 LST. The observed distribution and the Johnson S_B curve fit to that distribution are shown. The RMS between the observed distribution and modeled CDF is 0.5 percent, and the maximum difference is 0.8 percent.

5.3 Use of the Single-station Model

Suppose one wishes to find the probability of less than 0.50 sky cover at Moscow in January at 0600 LST. One would proceed as follows

For Moscow in January at 0600 LST, $\gamma = -\ 0.77442662$
$\eta = \ \ 0.12484847$

Equation (215) is used to calculate an END given a threshold sky cover

$$\ddot{z} = \gamma + \eta \cdot \ln[\ x_T/(1-x_T)\ ]$$

$$= -0.77442662 + 0.12484847 \ln[\ 0.5/(1.0-0.5)\ ]$$

$$= -0.77442662$$

Using a table of areas under a standard normal curve produces the required probability

$$\Pr(\ddot{z} \leq -0.77442662) = \Pr(X \leq 0.5) = 0.221$$

In the same manner, Equation (216) may be used to calculate a threshold sky cover given the value for the probability. One might want to know what threshold value of sky cover is exceeded 25 percent of the time at Kazan, RS, in June at 1500 LST.

For Kazan in June at 1500 LST, $\gamma = -\ 0.48898128$
$\eta = \ \ 0.43001788$

$\ddot{z}$ corresponding to the probability (that X is less than or equal to x_T) of 0.75 is 0.675. Equation (216) is used to calculate the threshold sky cover

$$x_T = \frac{1 + \exp[(\ddot{z}-\gamma)/\eta]}{\exp[(\ddot{z}-\gamma)/\eta]}$$

$$= \frac{\exp\ [(0.675 + 0.48898128)/0.43001788]}{1 + \exp\ [(0.675 + 0.48898128)/0.43001788]}$$

$$= 0.937$$

Using the model one could expect Kazan to have a sky cover greater than 0.937 25 percent of the time in June at 1500 LST (0.937 coverage converts to 8/8 when dealing with sky cover categories).

5.4 Modeling Temporal and Spatial Correlation of Sky Cover

5.4.1 Requirement for Correlation Matrices. In Chapter 4 a theorem from Anderson (1958) was presented for the generation of a random vector ($\underline{X}$). The joint sky cover probability model in this case study was based on the generation algorithm derived from Anderson's theorem

$$\underline{X} = \underline{C}\,\underline{\eta}$$

where $\underline{\eta}$ is a vector of random numbers distributed N(0,1) and $\underline{C}$ is a unique lower triangular matrix such that the correlation matrix $\underline{R}$ is equal to the product of $\underline{C}$ and the transpose of $\underline{C}$ (designated $\underline{C}'$). That is,

$$\underline{R} = \underline{C}\,\underline{C}'$$

One method for deriving $\underline{C}$ from $\underline{R}$ was presented in Chapter 4, namely, the Cholesky or "square-root" method. It is obvious that a good method of constructing the correlation matrix $\underline{R}$ is needed, since $\underline{R}$ is ultimately used to generate the vectors of ENDs of sky cover that produce the joint probability tables.

5.4.2 Spatial Correlation. Gringorten's Model-B (Gringorten, 1979) was used to model the spatial correlation function for this project. The Gringorten spatial correlation model is discussed in more detail in Chapter 6. Gringorten's equation for spatial correlation between two locations is

$$r = \frac{2}{\pi}\left[\ \cos^{-1}(\sigma) - \sigma\sqrt{1-\sigma^2}\ \right] \qquad (217)$$

where

$$\sigma \;=\; (\text{Actual Distance}) \,/\, (128 * \text{Scale Distance}) \qquad (218)$$

The scale distance is determined from observed data in the geographic area of interest. Gringorten's Model-B conforms to some preconceptions one has about a spatial correlation function. It is desirable that the function decrease exponentially with distance (not squared), at least for short distances, and drop to zero in a larger but finite distance. Table 24 compares the spatial correlation coefficients obtained from Gringorten's Model-B (Scale Distance = 7.8 km) with tetrachoric correlation coefficients calculated from observed data for various station pairs. It can be seen that although Model-B does not fit all cases, the fit for the overall data is not bad. Seasonal variations in the spatial correlation function can be accounted for by adjusting the scale distance. Table 25 lists the correlation coefficients as calculated by Model-B at various distances for different scale distances. This table should give a potential user a feel for how much the spatial correlation function can be altered by adjusting the scale distance.

Table 24. Spatial Correlation Coefficients Calculated by Gringorten's Model-B Compared to Tetrachoric Correlation Coefficients Computed from Observed Data.

	Spatial Correlation Coefficients			
	Station Pairs			
	Vladimar and Moscow	Voronezh and Moscow	Kingisepp and Moscow	Feddosiya and Moscow
Distance (km)	179	461	852	1199
Computed Cor. Coef.				
JAN	0.783	0.637	0.309	0.137
APR	0.694	0.120	0.184	0.012
JUL	0.526	0.335	0.160	0.239
OCT	0.717	0.425	0.209	0.039
ALL MONTHS	0.699	0.394	0.215	0.123
Modeled Cor. Coef.	0.771	0.430	0.215	0.000

Scale Distance for Gringorten Model = 7.8 km

Table 25. Spatial Correlation Coefficients from Gringorten's Model-B at Various Distances for Selected Scale Distances (The actual and scale distance must be in the same units).

	ACTUAL DISTANCE									
SCALE DISTANCE	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	450.0	500.0
1.0	0.516	0.119	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2.0	0.753	0.516	0.299	0.119	0.004	0.000	0.000	0.000	0.000	0.000
3.0	0.835	0.672	0.516	0.368	0.234	0.119	0.031	0.000	0.000	0.000
4.0	0.876	0.753	0.632	0.516	0.404	0.299	0.203	0.119	0.050	0.004
5.0	0.901	0.802	0.704	0.609	0.516	0.426	0.340	0.260	0.185	0.119
6.0	0.917	0.835	0.753	0.672	0.593	0.516	0.441	0.368	0.299	0.234
7.0	0.929	0.858	0.788	0.718	0.649	0.582	0.516	0.451	0.389	0.328
8.0	0.938	0.876	0.814	0.753	0.692	0.632	0.573	0.516	0.459	0.404
9.0	0.945	0.890	0.835	0.780	0.726	0.672	0.619	0.567	0.516	0.465
10.0	0.950	0.901	0.851	0.802	0.753	0.704	0.656	0.609	0.562	0.516

5.4.3 Serial or Temporal Correlation. For the temporal correlation, Equation (121) was used

$$r = 0.945^{\Delta t}$$

where r is the correlation coefficient between observations of sky cover at an initial time and some time lag (Δt in hours) and 0.945 is an empirical constant derived from observed data. Table 26 compares the temporal correlation coefficients calculated from Equation (115) to the tetrachoric correlation coefficients derived from observed data at selected stations for all hours, all months. As seen from Table 26, the correlation coefficients from Equation (121) are fairly close to those calculated from observed data in the first 18 hours but tend to approach zero faster than the observed coefficients beyond 18 hours. Once again, seasonal variations in the temporal correlation function can be accounted for by adjusting the constant.

Table 26. Temporal Correlation Coefficients Modeled from Gringorten's Equation Compared to Tetrachoric Correlation Coefficients Calculated from Observed Data.

	Temporal Correlation Coefficients			
	From Model		From Obsvd Data	
Time Lag		Kingisepp	Moscow	Chiganak
3	0.844	0.856	0.825	0.814
6	0.712	0.682	0.699	0.637
9	0.601	0.594	0.618	0.513
12	0.507	0.553	0.571	0.419
18	0.361	0.359	0.456	0.405
24	0.257	0.330	0.439	0.476
48	0.066	0.269	0.252	0.320

5.5 Models for the Joint Probability of Sky Cover

5.5.1 Joint Probability Models. In order to satisfy the two separate joint probability requirements of USAFETAC Project 2357, two operational models were developed. JSKY1 is the name for the USAFETAC model that produces joint sky cover distributions for a selected station at some designated time and N lag times (temporal problem), and JSKY2 is the USAFETAC model that produces joint sky cover distributions for selected station pairs (spatial problem). Both models are quite similar in the methods used to generate the joint probability of sky cover tables. The main difference is the technique for constructing the correlation matrix used to generate the vectors of correlated elements. In JSKY1 temporal correlation is used, while in JSKY2 spatial correlation is used. The models will now be examined in some detail.

5.5.2 JSKY1 Model. The overall plan of the model JSKY1 is shown in Figure 14. A precondition of using this model is that the Johnson S_B coefficients for the stations of interest must be available in a data file for call by the main program. The user provides the WMO station number of the particular location of interest and sets up a queue for the initial times, time lags, and the specific months for which the joint probability of sky cover tables are to be constructed. The final input parameter is the number of vectors that will be generated to construct each table. Each vector is independent of each other vector ($\rho = 0$). It is the elements within the vector that are correlated. If 3500 is specified, then the tables will contain an effective sample size of 3500 observations (see Equation 142).

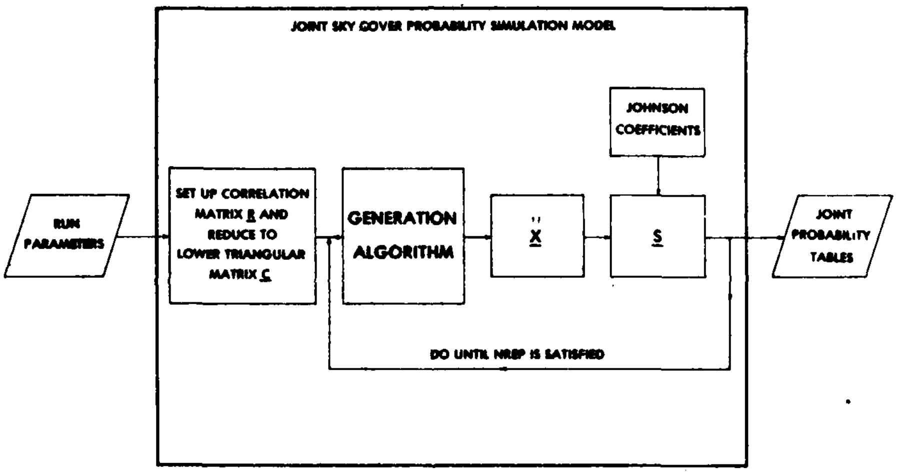

JOINT SKY COVER PROBABILITY SIMULATION MODEL
RUN PARAMETERS
SET UP CORRELATION MATRIX R AND REDUCE TO LOWER TRIANGULAR MATRIX C
GENERATION ALGORITHM
X
JOHNSON COEFFICIENTS
S
JOINT PROBABILITY TABLES
DO UNTIL NREP IS SATISFIED

Figure 14. Macro-design of the Joint Probability of Sky Cover Models Showing Flow of Information Through the Models.

Consider the following example. The problem is to compute the joint probabilities of sky cover for Kingisepp, RS, for September at 1200 GMT and lag times of 3, 6, 12, and 24 hours. Using Equation (121), the following correlation matrix is set up

Station: 260590 Initial Time: 1200 GMT

		Lag Time				
		0	3	6	12	24
Lag Time	0	1.00	0.84	0.71	0.51	0.26
	3	0.84	1.00	0.84	0.60	0.30
	6	0.71	0.84	1.00	0.71	0.36
	12	0.51	0.60	$\boxed{0.71}$	1.00	0.51
	24	0.26	0.30	0.36	0.51	1.00

For example, the 6- and 12-hour time lag observations are 6 hours apart ` thus are related by the expression,

$$r(\Delta t) = 0.945^6 = 0.71$$

The correlation matrix $\underline{R}$

$$\underline{R} = \begin{vmatrix} 1.00 & 0.84 & 0.71 & 0.51 & 0.26 \\ 0.84 & 1.00 & 0.84 & 0.60 & 0.30 \\ 0.71 & 0.84 & 1.00 & 0.71 & 0.36 \\ 0.51 & 0.60 & 0.71 & 1.00 & 0.51 \\ 0.26 & 0.30 & 0.36 & 0.51 & 1.00 \end{vmatrix}$$

is lower triangularized using USAFETAC subroutine LUSQRT, which implements the Cholesky decomposition scheme. The result is the lower triangular matrix $\underline{C}$,

$$\underline{C} = \begin{vmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.84 & 0.54 & 0.00 & 0.00 & 0.00 \\ 0.71 & 0.45 & 0.54 & 0.00 & 0.00 \\ 0.51 & 0.32 & 0.38 & 0.70 & 0.00 \\ 0.26 & 0.16 & 0.19 & 0.36 & 0.86 \end{vmatrix}$$

from which it can be verified that,

$$\underline{R} = \underline{C}\ \underline{C}'$$

Table 27. Steps in Generating a Random Vector of N Correlated Elements of Sky Cover.

1. Build a correlation matrix $\underline{R}$ using an appropriate correlation model (i.e., Gringorten's Model-B or the exponential decay model).

2. Obtain the lower triangular matrix $\underline{C}$ from $\underline{R}$ using USAFETAC subroutine LUSQRT (the Cholesky reduction scheme).

3. Generate N independent standard normal numbers.

$$(\eta_1, \eta_2, \ldots, \eta_N)$$

4. Perform the matrix-vector multiplication using the theorem from Anderson (1958).

$$\underline{X} = \underline{C} \cdot \underline{\eta}$$

5. Transform each of the elements of $\underline{X}$ into values of actual sky cover using an appropriate transnormalizing function (i.e., the Johnson S_B curve).

Table 27 summarizes the steps that are required to generate tables of joint sky cover probabilities. The lower triangular matrir $\underline{C}$ is passed to USAFETAC subroutine RANDCV. This subroutine generates a vector $\underline{\eta}$ of independent random standard normal numbers (i.e., numbers that are distributed N(0,1)). There are many random normal number generators that can be used. An example of this vector of independent numbers might be

$$\begin{aligned} \eta_1 &= -1.1006500 \\ \eta_2 &= 0.4851688 \\ \eta_3 &= -0.5071453 \\ \eta_4 &= -0.1079881 \\ \eta_5 &= -0.3342136 \end{aligned}$$

RANDCV then performs the matrix-vector multiplication specified by Anderson's theorem,

$$\underline{X} = \underline{C}\ \underline{\eta}$$

and the following vector $\underline{X}$ results

x_1 = -1.1006500

x_2 = -0.6685610

x_3 = -0.8362812

x_4 = -0.6713913

x_5 = -0.6285658

Since the elements of $\underline{X}$ are distributed N(0,1), they may represent correlated ENDs of sky cover. The Johnson S_B curve is then used to tranform these ENDs to sky cover categories by means of Equation (216) and the modeling coefficients for the time and month of the observation: (Note in the table below that the special term "oktas" refers to the number of eighths of sky cover.)

Vector Element		Sky Cover (Fractional Coverage)		Sky Cover Category (OKTAS)
S_1 (Time t)	=	0.2067654	=	2
S_2 (t + 3hr)	=	0.3663973	=	3
S_3 (t + 6hr)	=	0.0784657	=	1
S_4 (t + 12hr)	=	0.0369868	=	0
S_5 (t + 24hr)	=	0.4866438	=	5

Thus, the accumulator for sky cover (t) = 2, sky cover (Δt) = 3 is incremented for the 3-hour lag time table, sky cover (t) = 2, sky cover (Δt) = 1 for the 6-hour lag time table, etc., for all lag times. This procedure is repeated until the desired number of observations is achieved. An estimate of the joint probabilities is computed from the raw counts, and the simulation advances to the next hour or month until the hour and month queues are exhausted.

It should be emphasized here that the subroutine RANDCV produces a vector $\underline{X}$ of elements that are distributed multivariate normally according to the correlation specified in the correlation matrix $\underline{R}$. The questions remain, what type of degradation is involved in transforming each of the elements of $\underline{X}$ individually into sky cover categories by the Johnson S_B curve and how close to multivariate normality are the observed data?

Tables 28 and 29 compare the joint probability tables of observed and simulated sky cover data for Kingisepp, RS. The observed probabilities are based on observations from the 7-year period of record January 1973 through December 1979, and the simulated probabilites are based on 3500 synthetic observations. Table 28 contains the data for 0000 GMT, January, and a lag time of 12 hours, which represents a period in which the observed sky cover distributions were fitted to Johnson S_B curves with a great deal of success (i.e., low RMS values). Table 29 contains data for April, 0000 GMT, and a lag time of 6 hours, which represents a

Table 28. Observed and Simulated Joint Sky Cover Distributions for WMO Station 260590, Kingisepp, RS, at 0000 GMT, January and 12 Hour Lag Time.

OBSERVED

SKY COVER FOR INITIAL TIME (OKTAS)

Sky Cover for
Lag Time of 12 HRS
(OKTAS)

	0	1	2	3	4	5	6	7	8	TOT
0	0.089			0.009		0.005	0.005	0.005	0.061	0.174
1	0.009								0.019	0.028
2	0.009						0.005		0.028	0.042
3	0.014							0.005	0.028	0.047
4									0.005	0.005
5	0.005		0.005					0.005	0.005	0.019
6	0.019		0.009					0.009	0.052	0.089
7	0.019					0.005			0.078	0.094
8	0.061		0.005	0.014		0.005	0.009	0.014	0.394	0.502
TOT	0.225		0.019	0.023		0.014	0.019	0.038	0.662	1.000

SIMULATED

0	0.064	0.008	0.003	0.003	0.002	0.002	0.002	0.005	0.061	0.149
1	0.024	0.001	0.001	0.000	0.002	0.001	0.001	0.002	0.035	0.068
2	0.007	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.022	0.034
3	0.010	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.016	0.029
4	0.009	0.001	0.000	0.001	0.000	0.000	0.001	0.001	0.018	0.030
5	0.011	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.021	0.035
6	0.012	0.001	0.001	0.001	0.001	0.000	0.001	0.003	0.034	0.054
7	0.018	0.003	0.003	0.001	0.001	0.002	0.001	0.003	0.059	0.091
8	0.048	0.006	0.005	0.006	0.005	0.004	0.005	0.009	0.422	0.510
TOT	0.203	0.022	0.013	0.014	0.013	0.011	0.013	0.026	0.686	1.000

month in which the observed sky cover distributions were difficult to fit to the Johnson S_B curve (i.e., high RMS values). The largest difference between the observed and simulated tables for the January case is 2.8 percent. Considering the fact that the observed table is based on less than 250 observations, the values for the simulated table come well within the possible error intervals imposed from sampling theory alone. The largest difference between the observed and simulated joint probability tables for the April case occurs in the 8/8-8/8 joint occurrence category and is 5.2 percent. Even in this worst case month, the differences between the observed and simulated data are well within the limits expected from sampling theory, because of the small sample size.

5.5.3 JSKY2 Model. JSKY2 is the name of the USAFETAC model that produces joint sky cover distributions for selected station pairs at any desired time. Figure 14 illustrates the overall design of this model also.

Table 29. Observed and Simulated Joint Sky Cover Distributions for WMO Station 260590, Kingisepp, RS, at 0000 GMT, April, and 6-Hour Lag Time.

OBSERVED

SKY COVER FOR INITIAL TIME (OKTAS)

Sky Cover for
Lag Time of 6 HRS
(OKTAS)

	0	1	2	3	4	5	6	7	8	TOT
0	0.129		0.010	0.005		0.005	0.010		0.010	0.170
1	0.021			0.005				0.010	0.005	0.041
2	0.046		0.005			0.005	0.005	0.010	0.010	0.082
3	0.010								0.005	0.015
4									0.015	0.015
5	0.010		0.005				0.010	0.005		0.031
6	0.067		0.005						0.026	0.098
7	0.052	0.005		0.010		0.010	0.005	0.010	0.067	0.160
8	0.072		0.010			0.015	0.005	0.036	0.247	0.387
TOT	0.407	0.005	0.036	0.021		0.036	0.036	0.072	0.387	1.000

SIMULATED

0	0.127	0.006	0.003	0.001	0.002	0.001	0.002	0.002	0.008	0.152
1	0.055	0.003	0.004	0.001	0.002	0.000	0.001	0.005	0.010	0.082
2	0.029	0.005	0.001	0.001	0.002	0.001	0.001	0.002	0.007	0.050
3	0.020	0.003	0.001	0.001	0.001	0.001	0.001	0.001	0.006	0.035
4	0.021	0.005	0.001	0.001	0.002	0.001	0.002	0.002	0.009	0.044
5	0.023	0.003	0.002	0.001	0.001	0.003	0.001	0.004	0.014	0.052
6	0.022	0.003	0.002	0.001	0.002	0.003	0.002	0.005	0.020	0.060
7	0.037	0.010	0.004	0.005	0.003	0.002	0.004	0.009	0.038	0.110
8	0.041	0.018	0.007	0.008	0.008	0.006	0.010	0.018	0.299	0.416
TOT	0.374	0.056	0.025	0.021	0.023	0.017	0.024	0.049	0.412	1.000

As with model JSKY1, the Johnson S_B coefficients for possible stations of interest must be available in a data file for call by the main program. In addition to this requirement, a scale distance must be provided in order to tune the Gringorten spatial correlation function to the geographic area of interest. The user provides the WMO station numbers of the two locations to be modeled and then sets up a queue for the times and specific months that the tables of joint probability of sky cover are to be constructed for. The final input parameter is the number of vectors that will be generated to construct each table.

Consider the following example. The problem is to construct a table of joint probabilities of sky cover for Kazan, RS, and Vladimar, RS, in September at 1700 GMT. The distance between these two stations is calculated by the model as 548 km (see Chapter 6 for a detailed explanation of how great circle distance is calculated), and the Gringorten Model-B scale distance for sky cover is taken to be 7.8 km for that geographic area. Table 27 summarizes the steps necessary to generate the joint probability tables for the spatial problem. Equation (217) is used to calculate the spatial correlation coefficient for sky cover given the scale of the geographic area and the actual distance separating the two locations.

The following correlation matrix is set up

		Station #	
		275950	275320
Station #	275950	1.00	0.34
	275320	0.34	1.00

More stations could be added and the two-by-two matrix would be expanded to an N-by-N matrix. The correlation matrix $\underline{R}$,

$$\underline{R} = \begin{vmatrix} 1.00 & 0.34 \\ 0.34 & 1.00 \end{vmatrix}$$

is lower trangularized by USAFETAC subroutine LUSQRT,

$$\underline{C} = \begin{vmatrix} 1.00 & 0.00 \\ 0.34 & 0.94 \end{vmatrix}$$

and then USAFETAC subroutine RANDCV is used to generate the vectors of correlated elements distributed N(0,1) in the same manner as JSKY1. As indicated before, the main difference between the two models is in setting up the correlation matrix. After the matrix is set up, the two models proceed in the same way, by generating the vectors of ENDs and converting the vector elements to sky cover categories via the Johnson S_B curves.

Tables 30 and 31 compare the joint probability tables of observed and simulated sky cover for Moscow, RS, and Vladimir, RS. Once again the observed distributions are based on observations from the 7-year period of record January 1973 to December 1979, and the simulated probabilities are based on 3500 synthetic observations. Table 30 contains the data for 0000 GMT, January and represents a period in which the observed marginal distributions of sky cover were fitted to Johnson S_B curves with a great deal of success (i.e., low RMS values). The largest difference between the observed and simulated data in any joint occurrence category is 3 percent. The observed probabilities are based on a sample of size only 210 observations, so a difference of 3 percent falls well within the possible variations that might result from sampling error alone.

Table 30. Observed and Simulated Joint Sky Cover Distributions for Moscow and Vladimar, RS, at 0000 GMT, January.

OBSERVED

SKY COVER FOR MOSCOW (OKTAS)

SKY COVER FOR VLADIMAR (OKTAS)	0	1	2	3	4	5	6	7	8	TOT
0	0.124		0.024	0.005		0.014	0.038	0.014	0.029	0.248
1										
2	0.005							0.005	0.014	0.024
3	0.005					0.005		0.005		0.014
4					0.005				0.005	0.010
5	0.005			0.005					0.005	0.014
6	0.024						0.005		0.010	0.038
7	0.005			0.010	0.005		0.010	0.005	0.019	0.052
8	0.014		0.024				0.019	0.029	0.514	0.600
TOT	0.181		0.048	0.019	0.010	0.019	0.071	0.057	0.595	1.000

SIMULATED

0	0.110	0.020	0.008	0.006	0.007	0.007	0.008	0.014	0.039	0.220
1	0.008	0.001	0.001	0.000	0.002	0.002	0.001	0.003	0.008	0.027
2	0.002	0.002	0.000	0.000	0.001	0.001	0.001	0.002	0.007	0.016
3	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.004	0.009
4	0.003	0.002	0.000	0.000	0.000	0.001	0.002	0.002	0.006	0.015
5	0.002	0.001	0.001	0.001	0.000	0.000	0.001	0.002	0.007	0.015
6	0.004	0.001	0.000	0.002	0.001	0.000	0.001	0.001	0.007	0.017
7	0.004	0.003	0.003	0.001	0.002	0.002	0.004	0.004	0.018	0.040
8	0.019	0.013	0.010	0.006	0.008	0.008	0.014	0.039	0.522	0.640
TOT	0.153	0.045	0.024	0.017	0.021	0.021	0.032	0.068	0.618	1.000

Table 31 compares the joint probability tables of observed and simulated sky cover for these two stations for October, 1200 GMT. This time period represents a month in which the observed sky cover distributions were difficult to fit to the Johnson S_B curve (i.e., high RMS values). The largest difference between the observed and simulated data is only 4.8 percent (still within the size of possible variations that might be caused by sampling error alone).

5.6 Summary and Conclusions

The multivariate triangular matrix model MULTRI has been successfully used in operational multivariable, multistation models. The models were found to preserve the marginal distributions of sky cover at various Soviet locations as described by the Johnson S_B curve, which is guaranteed by the mathematics of the model. In addition, the models give a faithful representation of the joint probabilities of sky cover at station pairs or sky cover at a single location at an initial time and various time lags, which is not guaranteed by the mathematics of

Table 31. Observed and Simulated Joint Sky Cover Distributions for Moscow and Vladimar, RS, at 1200 GMT, October.

OBSERVED

SKY COVER FOR MOSCOW (OKTAS)

SKY COVER FOR VLADIMAR (OKTAS)

	0	1	2	3	4	5	6	7	8	TOT
0	0.024		0.019	0.005						0.048
1			0.010		0.005		0.005			0.019
2	0.005	0.005	0.010	0.005			0.010	0.005		0.038
3			0.005	0.005		0.005	0.005	0.029	0.010	0.057
4			0.005						0.005	0.010
5	0.005						0.005	0.005	0.019	0.033
6		0.005				0.019	0.019	0.024	0.029	0.095
7	0.005		0.014	0.005	0.005	0.014	0.029	0.071	0.095	0.238
8		0.005	0.005	0.005		0.010	0.019	0.029	0.390	0.462
TOT	0.038	0.014	0.067	0.024	0.010	0.048	0.090	0.162	0.548	1.000

SIMULATED

0	0.014	0.006	0.002	0.003	0.002	0.003	0.003	0.002	0.000	0.034
1	0.009	0.006	0.004	0.006	0.005	0.004	0.005	0.004	0.003	0.046
2	0.005	0.004	0.003	0.004	0.001	0.005	0.004	0.005	0.002	0.033
3	0.004	0.006	0.003	0.004	0.003	0.003	0.008	0.004	0.005	0.041
4	0.001	0.003	0.004	0.005	0.003	0.002	0.007	0.007	0.009	0.042
5	0.001	0.003	0.002	0.002	0.003	0.007	0.006	0.013	0.014	0.050
6	0.000	0.005	0.004	0.003	0.004	0.007	0.008	0.018	0.029	0.077
7	0.001	0.002	0.003	0.005	0.006	C.009	0.015	0.034	0.078	0.153
8	0.000	0.002	0.003	0.001	0.005	0.006	0.014	0.054	0.438	0.523
TOT	0.036	0.036	0.027	0.032	0.033	0.044	0.071	0.141	0.579	1.000

the model. Furthermore, Gringorten's models for the spatial and temporal correlation functions of sky cover were tested and were found to give very good estimates of the actual correlation functions calculated from observed data.

Chapter 6

A MODEL FOR THE SIMULATION OF GRIDDED FIELDS

6.1 General

The multivariate triangular matrix model MULTRI, discussed in Chapters 4 and 5, represents a very powerful and flexible tool that USAFETAC uses in its environmental simulation applications. This technique allows a cross correlated vector of N weather variables to be produced at the request of the user. At times it becomes necessary for USAFETAC to simulate two-dimensional gridded fields of meteorological variables that are correlated in space. It is for these applications that the MULTRI model is usually inappropiate. The triangular matrix model becomes cumbersome from a mathematical/computational point of view whenever the number of variables N exceeds 30. In actual practice, USAFETAC prefers to limit the number of correlated variables to 15 or less. A small 10 x 10 gridded field would require 100 cross correlated elements, so even for such a small problem as that, another approach would have to be used.

One such technique is the two-dimensional field simulation model (2DFLD), which is based on a sawtooth wave submodel developed by Maj Albert Boehm, USAFETAC/DNP, for USAFETAC Project 1960, Colossus Weather Simulation. The sawtooth wave model is used to generate a two-dimensional, spatially correlated field of random normal numbers. As in the multivariate triangular matrix model, these random normal numbers distributed N(0,1) may represent ENDs of some weather variable and then be individually transformed by some transnormalizing function to raw values of that variable. In this chapter the basic design of the sawtooth wave generator will be discussed. The reader should consult USAFETAC/TN-81/004, Cloud Forecast Simulation Model (Whiton, et al., 1981), for an example of an application of this technique to an operational problem.

6.2 Spatial Correlation Function of the Random Normal Number Field

The sawtooth wave submodel generates a field of ENDs η having a desired spatial correlation function r. Consider the correlation r between values η_j and $\eta_{j+\Delta t}$ located one grid distance Δj apart. This is shown in Figure 15. Repeated samplings of the value of η at j and at j+Δj would produce a history of N data pairs from which the spatial correlation could be estimated by the Pearson product moment formula,

$$r = \frac{\frac{1}{N}\sum_{k=1}^{N} \eta_{j,k}\,\eta_{j+\Delta j,k} - \overline{\eta_j}\;\overline{\eta_{j+\Delta j}}}{s_{\eta_j}\, s_{\eta_{j+\Delta j}}} \qquad (219)$$

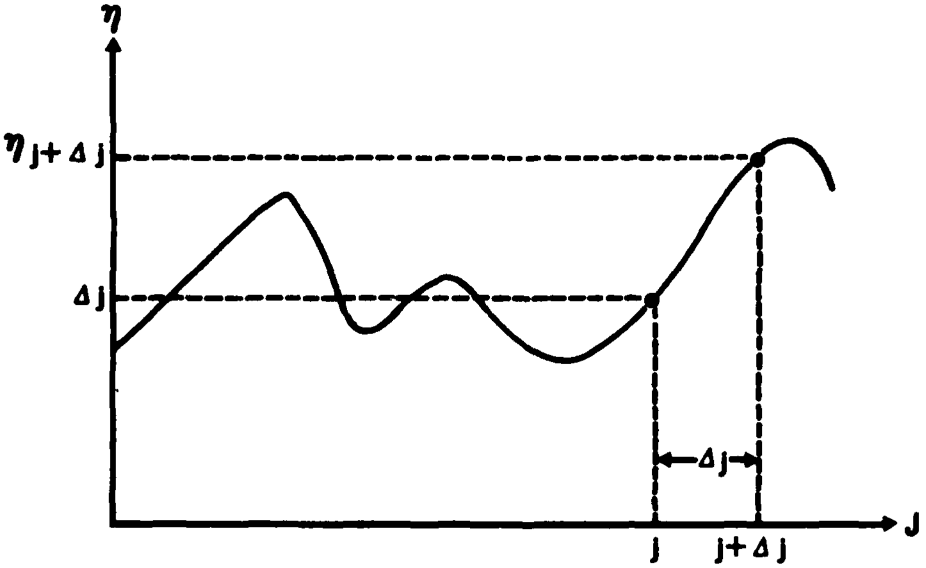


Figure 15. Correlation Between η at Location $J = j$ and η at Location $J = j + \Delta j$ Located One Grid Distance Δj Apart.

or by some other method. In Equation (219), the overbars represent means, and s represents the standard deviation. Since the η are ENDs, they are distributed normally with a mean of zero and a variance of unity. Therefore, for normally distributed η, Equation (219) reduces to

$$r = \frac{1}{N} \sum_{k=1}^{N} \eta_{j,k}\, \eta_{j+\Delta j,k} \qquad (220)$$

Spatial corrrelation is being dealt with here. The correlation between η values will be perfect (unity) at zero separation ($\Delta j = 0$) and will be less than or equal to unity with increasing distance Δj. To model the weather, a correlation function is needed that starts at unity and decreases with increasing distance d.

One such model that has been used successfully in ceiling, visibility and sky cover modeling is that of Gringorten (1979). In Gringorten's Model-B, the correlation function r depends on the geometric distance d and a characteristic scale distance D, defined as the distance at which the correlation r falls to 0.99. Gringorten's Model-B was presented in Chapter 5 as Equation (217) and is defined by,

$$r = r(d,D) = \frac{2}{\pi} [\cos^{-1}\sigma - \sqrt{1-\sigma^2}\,] \quad \text{(dimensionless)} \qquad (217)$$

recalling that,

$$\sigma = d/(128\ D) \qquad \text{(dimensionless)} \qquad (218)$$

Because $\sigma \geq 0$, the trigonometric relationship between arc cosine and arc sine can be used, yielding the form,

$$r = r(d,D) = \frac{2}{\pi}\ (\sin^{-1}H - \sigma H) \qquad (221)$$

where

$$H = \sqrt{1-\sigma^2} = \sqrt{1 - d^2/(16384\ D^2)} \qquad (222)$$

In this correlation function model, when the distance d equals the scale distance D, $\sigma = 1/128$, $H = 0.99997 \approx 1$, $\sin^{-1}(1) = \pi/2$, $\sigma H = 0.00781$, and $r = 0.99$. Note that when $\sigma = 1$, $H = 0$, and $r = 0$. Therefore, Gringorten's Model-B correlation drops to zero at a distance $d = 128\ D$. Gringorten has estimated the scale distance for sky cover in Germany as 4 km. Using this for D in Equations (221) and (222) gives the correlation function shown in Figure 16. With this scale distance, the correlation drops to 0.99 in 4 km (2 NM) and to approximately zero at 512 km (276 NM).

It is desired that the sawtooth wave model produce a field of ENDs having the spatial correlation function of Gringorten's Model-B, discussed above. During his earlier work with the sawtooth wave generator, Boehm developed empirical equaticns that converted a desired Gringorten Model-B scale distance (SD) into maximum and minimum allowable wavelengths (the interval from which the wavelengths of each sawtooth wave will be selected at random) for the sawtooth wave generator. These equations are

$$W\ (\text{upper}) \quad = \quad C(u) * SD \qquad (223)$$

$$W\ (\text{lower}) \quad = \quad C(l) * SD \qquad (224)$$

The dimensionless constants C(u) and C(l) are selected so the model returns a spatial correlation function for the random normal number field that has a shape similar to a Gringorten curve for the particular SD. The wavelengths will be in the same units as the scale distance. The values of 175 and 450 are currently used for C(l) and C(u), respectively.

6.3 The Mathematics of the Sawtooth Wave

In the sawtooth wave model, N sawtooth waves are allowed to emanate circularly from N focal points. Each focal point is the source of exactly one wave. The location of each focal point is picked at random, and the wavelength of each wave is selected at random from a range of permissible wavelengths. The field of ENDs

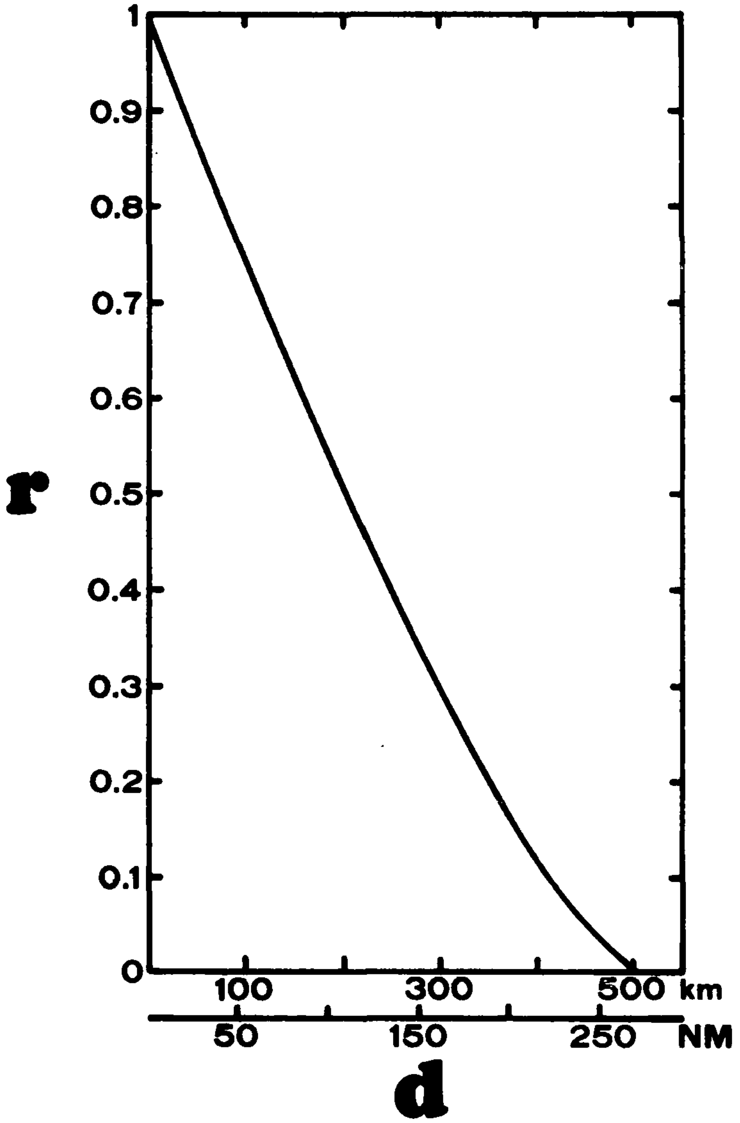


Figure 16. Correlation Function for Gringorten's Model B with Scale Distance D = 4 km.

η is simply the sum of N sawtooth wave amplitudes at each grid point, corrected by subtraction of a constant.

Each sawtooth wave is as shown in Figure 17. Amplitude of the wave, shown by y, varies between zero and one, depending on the observer's position along the wave. The sawtooth wave used here is a standing wave. Originating with zero amplitude at a focal point at distance d' = 0, it reaches maximum amplitude (unity) at distance d' = 1 wavelength, and thereafter falls to zero amplitude again. Within any one cycle of the sawtooth wave, the slope of wave amplitude versus distance is unity, i.e.,

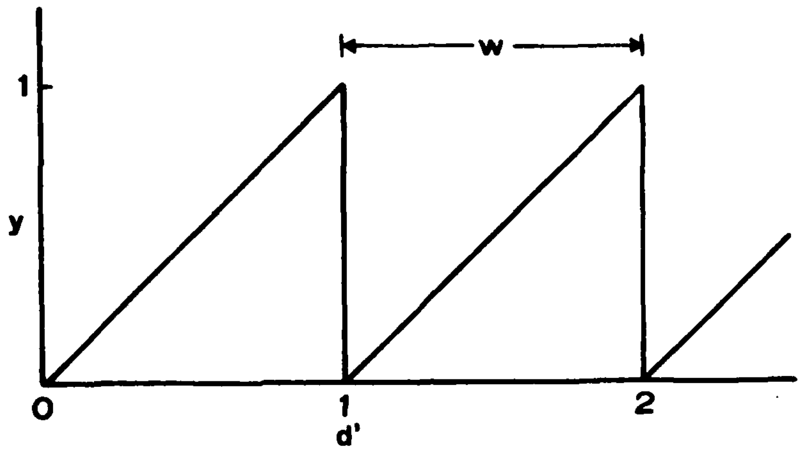


Figure 17. Sawtooth Wave. d' represents normalized distance orthogonal to the wave front. The normalized distance d' is measured in unit wavelengths, where wavelength is represented by w.

$$dy/dd' = 1 \tag{225}$$

Hence, within any one cycle of the sawtooth wave, its equation is

$$y = d' \tag{226}$$

To allow for multiple cycles of the sawtooth, it is appropriate to write its equation as

$$y = d' - \mathrm{INT}(d') \tag{227}$$

where INT(d') represents the largest integer less than or equal to the normalized distance d'. The normalized distance d' is the geometric distance d expressed in unit wavelengths w, i.e.,

$$d' = d / w \tag{228}$$

Hence,

$$y = d/w - \mathrm{INT}(d/w) \tag{229}$$

An alternative Fourier representation of the sawtooth wave is

$$y = \pi - 2 \sum_{i-1}^{\infty} \frac{\sin(\, id' \,)}{i} \qquad (230)$$

The simple form of the sawtooth wave makes it easy to calculate the amplitude y_{jk} of a wave at location j whose origin is the focal point at location k. This is done by computing the great circle distance between locations j and k, i.e.,

$$d = GCD(j,k) \qquad (231)$$

and then evaluating Equation (229) with the wavelength w known.

But any single wave amplitude y_{jk} does not create randomness. The η field produced by the sawtooth wave model must be random. Its elements η_j must have been selected at random from a normally distributed population with a mean of zero and a variance of one, i.e., N(0,1). It is apparent that the distribution of any one sawtooth wave is uniform, with a mean of 1/2 and a variance of 1/12. But the sum of approximately 12 uniform random numbers, by the central limit theorem, approaches the normal distribution. Naylor, et al. (1966) give the equation for calculating a normally distributed pseudorandom number G from the sum of N uniform pseudorandom numbers U

$$G = \sigma_G \sqrt{12/N} \; (\sum_{n=1}^{N} U_n - \frac{N}{2}) + \mu_G \qquad (232)$$

where σ_G and μ_G are the desired standard deviation and mean, respectively, of G. For the special case where $\sigma_G = 1$ and $\mu_G = 0$, and where the number of uniform random numbers to be summed is N = 12, Equation (232) simplifies to

$$G = \sum_{n=1}^{12} U_n - 6 \qquad (233)$$

Applied to the question of calculating a normally distributed value η_j for location j from the sum of N = 12 uniformly distributed sawtooth wave amplitudes y_{jk}, Equation (233) becomes

$$G = \sum_{k=1}^{12} y_{jk} - 6 \qquad (234)$$

The superposition of sawtooth waves is illustrated in Figure 18. Two sawtooth waves are shown emanating from randomly positioned focal points k = 1 and k = 2. These waves are converging on location j with respective amplitudes y_{j1} and y_{j2}. The wavelengths w_k of the two waves are illustrated as being different to emphasize that those wavelengths were drawn at random uniformly from a range of possible wavelengths.

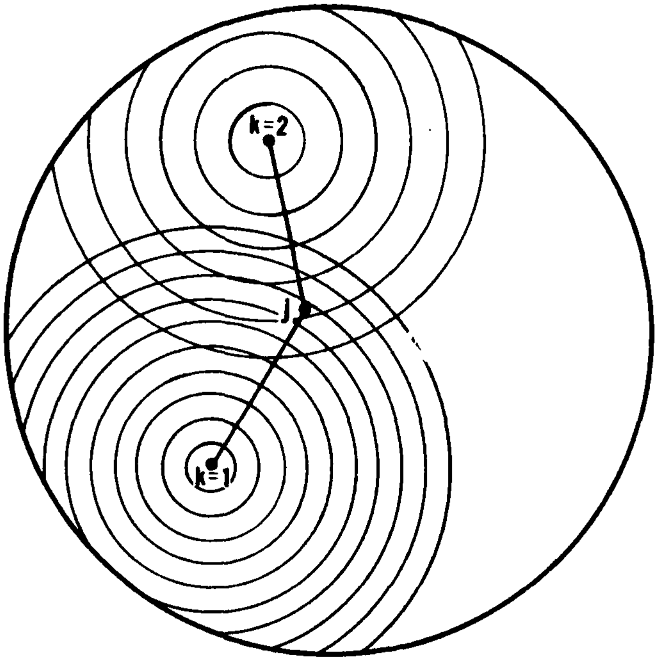


Figure 18. Sawtooth Waves Emanating from Focal Points at Locations k Converge on Location j.

6.4 Calculation of Great Circle Distance

The great circle distance d between any two points "a" and "b" on the globe can be calculated from the latitude and longitude of point "a" (θ_a, λ_a) and that of point "b" (θ_b, λ_b). The conventional equation is

$$d = r \cos^{-1}[\sin\theta_a \sin\theta_b + \cos\theta_a \cos\theta_b \cos(\lambda_a - \lambda_b)] \qquad (235)$$

where r is the radius of the earth, approximately 6371 km. This equation involves calculating five sines and cosines plus one arc cosine.

An alternative expression for the great circle distance d can be obtained by using the trigonometric function-product relations,

$$\sin\theta_a \sin\theta_b = (\tfrac{1}{2})[\cos(\theta_a - \theta_b) - \cos(\theta_a + \theta_b)] \qquad (236)$$

$$\cos\theta_a \cos\theta_b = (\tfrac{1}{2})[\cos(\theta_a - \theta_b) + \cos(\theta_a + \theta_b)] \qquad (237)$$

or

$$\sin\theta_a \sin\theta_b = (\tfrac{1}{2})(d - s) \qquad (238)$$

$$\cos\theta_a \cos\theta_b = (\tfrac{1}{2})(d + s) \qquad (239)$$

where

$$d = \cos(\theta_a - \theta_b) \tag{240}$$

$$s = \cos(\theta_a + \theta_b) \tag{241}$$

from which it is found that

$$d = r \cos^{-1}\{(½)[(d - s) + (d + s)\cos(\lambda_a - \lambda_b)]\} \tag{242}$$

This equation calls for evaluating three cosines and one arc cosine and should therefore be much faster to solve than Equation (235).

A third expression for the great circle distance can be obtained by using the trigonometric angle-difference relation,

$$\cos (\lambda_a - \lambda_b) = \cos\lambda_a \cos\lambda_b + \sin\lambda_a \sin\lambda_b \tag{243}$$

in Equation (235), from which it is found that

$$d = r \cos^{-1}[\sin\theta_a \sin\theta_b + \cos\theta_a \cos\theta_b(\cos\lambda_a \cos\lambda_b + \sin\lambda_a \sin\lambda_b)] \tag{244}$$

Because this equation involves eight sines and cosines plus one arc cosine, it appears at first glance much less suitable for use than Equations (235) or (242). Nevertheless, Equation (244) offers some "operational" advantages that make it useful. In particular, one need not know the actual latitudes and longitudes to calculate great circle distance from Equation (244); only the sines and cosines of the latitudes and longitudes are needed. Moreover, since the number of focal points is small (generally 12 or fewer), the needed sines and cosines can be calculated initially and then stored for repeated use.

6.5 Selection of Focal Points and Wavelengths for the Sawtooth Wave

6.5.1 Selection of Focal Point. Each sawtooth wave must emanate from a randomly positioned focal point; otherwise, the amplitude sums will not be random. Focal points are located in terms of their latitude θ_k and longitude λ_k, where, for convenience, the longitude ranges from 0 degrees through 360 degrees. The longitude λ_k of the kth focal point is selected uniformly from the range 0 degrees to 360 degrees by the equation,

$$\lambda_k = 360 U_k \tag{245}$$

where U_k is a pseudorandom number selected from a population uniformly distributed over the range (0,1).

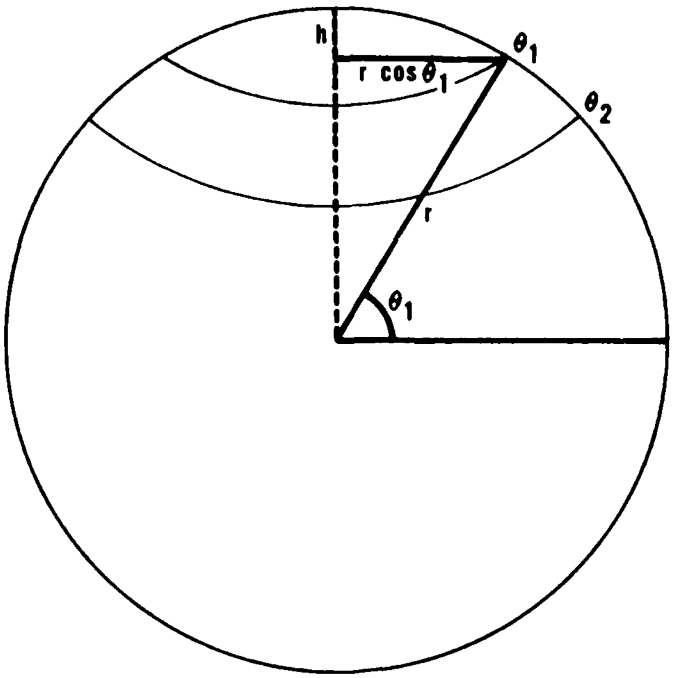


Figure 19. Geometry for Surface Area of the Spherical Zone Bounded by Latitudes θ_1 and θ_2, Where $\theta_1 > \theta_2$.

While the longitude λ_k of the focal point can be selected uniformly from the range 0 degrees to 360 degrees, it is not true that the latitude θ_k can be selected uniformly from the range 0 degrees to 180 degrees (90 degrees to -90 degrees). This is because equiprobable latitude bands are not equal area bands, and simple selection of latitude would result in an overly dense concentration of focal points per unit surface area near the poles. Figure 19 shows the geometry of this problem. Needed is an expression for the surface area of the spherical zone bounded by latitudes θ_1 and θ_2. The essential principle is that the surface area of the zone is the difference between the surface area of the spherical cap formed by θ_2 and that formed by θ_1.

Consider only the spherical cap formed by θ_1. This has height h in a sphere of radius r. The surface area of that cap is

$$S_1 = 2\pi rh \qquad (246)$$

But from the Pythagorean theorem,

$$x^2 = r^2 - r^2\cos^2\theta_1 = r^2(1 - \cos^2\theta_1) = r^2\sin^2\theta_1 \qquad (247)$$

and

$$x = r \sin\theta_1 \tag{248}$$

Moreover,

$$h = r - x \tag{249}$$

$$h = r(1 - \sin\theta_1) \tag{250}$$

Hence, the surface area of the spherical cap formed by θ_1 is, from Equations (246) and (250),

$$S_1 = 2\pi r^2(1 - \sin\theta_1) \tag{251}$$

By analogy, the surface area of the spherical cap formed by θ_2 is

$$S_2 = 2\pi r^2(1 - \sin\theta_2) \tag{252}$$

The surface area S_z of the zone is the difference,

$$S_z = S_2 - S_1 \tag{253}$$

$$S_z = 2\pi r^2(\sin\theta_1 - \sin\theta_2) \tag{254}$$

The function difference relations give the result,

$$\sin\theta_1 - \sin\theta_2 = 2 \cos \tfrac{1}{2}(\theta_1 + \theta_2) \sin \tfrac{1}{2}(\theta_1 - \theta_2) \tag{255}$$

Consider that the width of the latitude band θ_1 to θ_2 will always be constant, say 5 degrees or 10 degrees. Then the sine of one-half their difference is also a constant, say D:

$$\sin \tfrac{1}{2}(\theta_1 - \theta_2) = D \tag{256}$$

Thus,

$$\sin\theta_1 - \sin\theta_2 = 2\, D \cos \tfrac{1}{2}(\theta_1 + \theta_2) \tag{257}$$

Using Equation (257) in Equation (254) produces the result,

$$S_z = 4\pi D\, r^2 \cos \tfrac{1}{2}(\theta_1 + \theta_2) \tag{258}$$

But r is constant, so

$$C = 4\pi Dr^2 = \text{const} \tag{259}$$

and

$$\bar{\theta} = \tfrac{1}{2}(\theta_1 + \theta_2) \tag{260}$$

where $\bar{\theta}$ is the mean latitude of the zone bounded by θ_1 and θ_2. Using Equations (259) and (260) in Equation (258) produces the result,

$$S_z = C \cos \bar{\theta} \tag{261}$$

Equation (261) shows that the surface area of the spherical zone bounded by latitudes θ_1 and θ_2 is proportional to the cosine of the mean latitude of the zone. If we simply choose the latitude of the focal point uniformly over the permitted range of latitudes, then the density of selections will not show a poleward decrease proportional to the poleward decrease of zonal surface area S_z. This can be compensated for a selecting cos θ_k rather than θ_k itself. Since the cosine has the range [0,1], the equation is

$$\cos \theta_k = U_k' \tag{262}$$

where U_k' is a uniform pseudorandom number drawn from the same range. Selection of the latitude of the focal point in this way restricts the focal point to the Northern Hemisphere, but this imposes no limits on the randomness of the result.

6.5.2 Selection of Wavelengths. If the wavelength w_k of the sawtooth wave emanating from location k is to be selected from the interval,

$$w_1 < w_k \leq w_2 \tag{263}$$

such that any value is equally likely to be chosen, then the selection can be made by drawing a random number uniformly from Equation (263). If U_k'' is a pseudorandom number drawn from a uniform distribution having the range 0 to 1, then

$$w_k = U_k''(w_2 - w_1) + w_1 \tag{264}$$

An algorithmic procedure for the sawtooth wave generator is shown in Figure 20.

```
procedure SAWTOO (m, ETA);
integer j, k, m, n;
real w, d, y
real array YSUM [1:m], ETA[1:m];
equivalence (YSUM, ETA);
for each location or grid point j:  = 1 step 1 until m do
begin
    initialize $YSUM_j$: = 0.0;
end j;
for each focal point k:  = 1 step 1 until n do
comment:  ... n + 12 ...;
    select location of $k^{th}$ focal point at random;
        comment:  ... Equations (239) and (256) ...;
    select wavelength w at random from $(w_1,w_2)$;
        comment:  ... Equation (258) ...;
    for each location or grid point j:  = 1 step 1 until
    m do begin
        calculate distance d:  = GCD(j,k);
            comment:  ... Equation (238)
        calculate wave amplitude y;
            comment:  ... Equation (223) ...;
        accumulate $YSUM_j$ = $YSUM_j$ + y;
    end j;
end k;
for each location or grid point j:  = 1 step 1 until m do
begin
    $ETA_j$: = $YSUM_j$ - 6;
        comment:  ... Equation (228) ...;
end j;
end SAWTOO;
```

Figure 20. Algorithm for Sawtooth Wave Submodel.

6.6 Summary and Conclusions

The development of the two-dimensional field simulation model 2DFLD lays much of the groundwork for the solution of a class problem, the simulation of two-dimensional fields of a desired variable, fields that are correlated in space. Given a suitable transnormalizing function for the desired variable, the 2DFLD model can be used to produce random END fields with spatial correlations similar to those found in real data, and then the inverse normalizing function can be used to obtain synthetic fields of the variable.

REFERENCES AND BIBLIOGRAPHY

Acton, F.S., 1970: Numerical Methods That Work, (New York: Harper and Row, Publishers), p 340.

Air Weather Service, 1977: Guide for Applied Climatology, AWS-TR-77-267, p 4-31.

Anderson, T.W., 1958: An Introduction to Multivariate Statistical Analysis, (New York: John Wiley and Sons), p 19.

Bean, S.J., P.N. Somerville, and M. Heuser, 1979: Some Models for Ceiling, University of Central Florida, Scientific Report No. 7, Air Force Geophysics Laboratory, AFGL-TR-79-0221, 35pp.

Boehm, A.R., 1976: Transnormalized Regression Probability, Air Weather Service, AWS-TR-75-259, 52pp.

Box, G.E.P., and G.M. Jenkins, 1976: Time Series Analysis: Forecasting and Control, (San Francisco: Holden-Day, Inc.), 575pp.

Carnahan, B., H.A. Luther, and J.O. Wilkes, 1969: Applied Numerical Methods, (New York: John Wiley and Sons, Inc.), p 334.

Elderton, W. P., (1953): Frequency Curves and Correlation, (Washington DC: Harren Press), pp 141-180.

Forsythe, G.E., M.A. Malcolm, and C.B. Moler, 1977: Computer Methods for Mathematical Computations, (Englewood Cliffs, NJ: Prentice-Hall), 259pp.

Forsythe, G.E. and C.B. Moler (1967): Computer Solution of Linear Algebraic Systems, (Englewood Cliffs: Prentice-Hall, Inc.), pp 114-115.

Friend, A.L., 1978: An Objective Technique for Spreading Climatology, USAFETAC-PR-78-007, United States Air Force Environmental Technical Applications Center, 4pp.

Gringorten, I.I., 1979: Probability models of weather conditions occupying a line or area, J. Appl. Met., 18, pp 957-977.

Haan, C.T., 1977: Statistical Methods in Hydrology, (Ames: The Iowa State University Press), 378pp.

Heuser, M., P.N. Somerville, S.J. Bean, 1980: Least Squares Fitting of Distributions Using Non-Linear Regression, Air Force Geophysics Laboratory, AFGL-TR-80-0362, 18pp.

Hicks, P., 1982: Project notes. Unpublished USAFETAC manuscript.

Huschke, R.E., and R.R. Rapp, 1970: Weather Service Contribution to STRICOM Operations--A Survey, A Model and Results: Final Report on Phase I of the Rand Corporation Contribution to the Air Weather Service Mission Analysis, R-542-PR, The Rand Corporation, 58pp.

James, R.C., and G. James, 1968: Mathematics Dictionary, 3rd Ed., (New York: Van Nostrand Reinhold Company), pp 517.

Lin, C.C., and L.A. Segel, 1974: Mathematics Applied to Deterministic Problems in the Natural Sciences, (New York: MacMillan Publishing Co., Inc.), 604pp.

Loucks, D.P., J.R. Stedinger, and D.A. Haith, 1981: Water Resource Systems Planning and Analysis, (Englewood Cliffs: Prentice-Hall, Inc.), 559pp.

Lowry, G.G., (1970): Markov Chains and Monte Carlo Calculations in Polymer Sciences, (New York: Marcel Dekker, Inc.), pp 13-43.

Miller, R.G., and R.C. Whiton, 1979: A Weather Simulation Model Based on REEP, Preprints, 6th AMS Conference on Probability and Statistics in Atmospheric Sciences, American Meteorological Society, pp 167-172.

Naylor, T.H., J.L. Balintfy, D.S. Burdick and K. Chu, 1966: Computer Simulation Techniques, (New York: John Wiley and Sons, Inc.), pp 68-121.

Panofsky, H.A., and G.W. Brier, 1965: Some Applications of Statistics to Meteorology, (University Park, PA: The Pennsylvanis State University), 224pp.

Parzen, E., 1962: Stochastic Processes, (San Francisco: Holden-Day, Inc.), 324pp.

Pratte, J.F., and R.W. Lee, 1979: A short method of generating meteorological fields for simulation studies, J. Appl. Met., 18, pp 1670-1673.

Scheuer, F., and D.S. Stoller, (1962): On the generation of normal random vectors, Techometrics, IV, 278-281.

Somerville, P.N., and S.J. Bean, 1979: A New Model for Sky Cover, Air Force Geophysics Laboratory, AFGL-TR-79-0219, 33pp.

Somerville, P.N., S.J. Bean, and S. Falls, 1979: Some Models for Visibility. University of Central Florida, Scientific Report No. 3, Air Force Geophysics Laboratory, AFGL-TR-79-0144, 40pp.

Somerville, P.N., S. Watkins, and R. Daley, 1978: Some Models for Sky Cover. Florida Technological University, Scientific Report No. 2, Air Force Geophysics Laboratory, AFGL-TR-78-0219, 22pp.

Tatsuoka, M.M., 1971: Multivariate Analysis: Techniques for Educational and Psychological Research, (New York: John Wiley and Sons, Inc.), 310pp.

Whiton, R.C., E.M. Berecek, and J.G. Sladen, 1981: Cloud Forecast Simulation Model, USAFETAC/TN-81/004, United States Air Force Environmental Technical Applications Center, 134pp (AD-A113140).

Appendix A

SOME FUNCTIONS USED BY USAFETAC TO MODEL THE CUMULATIVE FREQUENCY DISTRIBUTIONS OF METEOROLOGICAL VARIABLES

Index of Functions

<u>Variable to be Modeled</u>. Sky Cover

<u>Function Name</u>. Johnson S_B Curve

<u>General</u>. USAFETAC's basic modeling equation for sky cover is the Johnson S_B family of curves. This function was first developed for fitting sky cover distributions by Somerville, Watkins, and Daley (1978). The Johnson S_B curve is given by the equation

$$\ddot{z} = \gamma + \eta \ln[\frac{x_T}{1-x_T}] \qquad \text{(A-1)}$$

where γ and η are modeling coefficients determined from empirical distributions, x_T is some threshold sky cover in fractional coverage, and $\ddot{z}$ is the equivalent normal deviate (END) of the cumulative frequency that the actual sky cover (X) is less than x_T. That is, this probability is the intergral of the standard normal distribution from $-\infty$ to $\ddot{z}$. This can be expressed by the equation

$$\Pr(X \leq x_T) = \Phi(x_T) = \int_{-\infty}^{\ddot{z}} \phi(u)\, du \qquad \text{(A-2)}$$

where $\phi(u)$ is the standard normal probability density function,

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \qquad \text{(A-3)}$$

The normal density function cannot be integrated directly, but there are many rational approximations or look-up tables available for evaluating this integral.

Inverse Transnormalizing. Values of sky cover can be obtained from given probabilities by solving Equation (A-1) for x_T

$$x_T = \frac{1 + e^{(\frac{\ddot{z}-\gamma}{\eta})}}{e^{(\frac{\ddot{z}-\gamma}{\eta})}} \qquad \text{(A-4)}$$

Fitting the Curve. The modeling coefficients γ and η can be obtained for any observed cumulative frequency distribution of sky cover by any good linear regression technique. USAFETAC uses the method of singular value decomposition described by Forsythe, et al., (1977). The values of $\ddot{z}$ corresponding to the percentage of time that the sky cover is less than some category is regressed against the interior boundary value of that category (x_T). For example, when the sky cover is observed in oktas, there are nine categories designated as 0, 1, 2, 3, 4, 5, 6, 7, and 8. The interior boundaries of these categories when expressed in decimal form are taken to be 0.0625, 0.1875, 0.3125, 0.4375, 0.5625, 0.6875, 0.8125, and 0.9375.

The values of γ and η are obtained using the singular value decomposition scheme to regress values of $\ddot{z}$ on x_T. The values of $\ddot{z}$ used are those corresponding to the END of the tabulated proportion of sky cover less than some category and the value of x_T is the interior boundary of that category. For example, the sky cover distribution at Vyborg, RS, for March, 2100 LST is

Sky Cover x_T	0	1	2	3	4	5	6	7	8
Obsvd Freq	24.6	1.0	7.7	2.9	1.0	1.4	5.3	3.9	52.2

Using these relative frequencies to compute the cumulative frequency that X is less than x_T, the table becomes

Sky Cover x_T	0	1	2	3	4	5	6	7	8
Cum Freq X < x_T	0	24.6	25.6	33.3	36.2	37.2	37.2	43.9	47.8

The ENDs of the cumulative distributions are then fitted against the interior boundaries of the categories.

Independent Variable (x_T)	Dependent Variable (P) * 100	END of P (Z)
0.0625	24.6	-0.6868
0.1875	25.6	-0.6554
0.3125	33.3	-0.4312
0.4375	36.2	-0.3527
0.5625	37.2	-0.3261
0.6875	38.6	-0.2893
0.8125	43.9	-0.1532
0.9375	47.8	-0.0550

Variable to be Modeled. Sky Cover

Function Name. S-Distribution

General. The S-distribution was first developed for fitting sky cover data by Somerville and Bean (1979). The S-distribution offers one advantage over the Johnson S_B, curve in that it has a closed form distribution. The probabilities can be obtained by direct substitution and no numerical integration or approximations are required. The S-distribution for sky cover is given by,

$$P = 1 - (1-x_T{}^{\alpha})^{\beta} \qquad \text{(A-5)}$$

where α and β are modeling coefficients determined from empirical distributions, x_T is some threshold sky cover in fractional coverage, and P is the probability that the actual sky cover (X) is less than or equal to x_T.

Inverse Transnormalizing. Values of sky cover can be obtained from given probabilities by solving Equation (A-5) for x_T

$$x_T = [\ 1 - (1-P)^{1/\beta}\ ]^{1/\alpha} \qquad \text{(A-6)}$$

Fitting the Curve. To obtain the modeling coefficients α and β for any observed cumulative frequency distribution of sky cover, nonlinear regression techniques must be used. USAFETAC uses an iterative search technique developed by Capt Robert Hughes, USAFETAC/DNB. The probabilities that the sky cover are less than some category are regressed against the interior boundaries for the category as described in the section of this appendix on the Johnson SB curve. The curve fitting technique used for this function is CPU intensive as compared to the linear methods for the Johnson S_B curves and comparative tests show that RMS values for fits of S-distribtuions are not superior to those of Johnson S_B curves, For these reasons USAFETAC uses the Johnson curve to model sky cover predominantly.

Variable to be Modeled. Visibility

Function Name. Weibull Curve

General. USAFETAC's basic modeling equation for visibility is the Weibull curve. The function was first applied to fitting visibility distributions by Somerville, Bean, and Falls (1979). The Weibull curve is given by the equation,

$$P = 1 - \exp(-\alpha x_T{}^{\beta}) \qquad \text{(A-7)}$$

where α and β are modeling coefficients determined from empirical distributions, x_T is some threshold visibility in statute miles, and P is the probability that the actual visibility (X) is less than or equal to x_T.

Inverse Transnormalizing. Values of visibility can be obtained from given probabilities by solving Equation (A-7) for x_T

$$x_T = \left[\frac{\ln(1-P)}{-\alpha} \right]^{1/\beta} \qquad \text{(A-8)}$$

Fitting the Curve. To obtain the modeling coefficients α and β, the values for an empirical cumulative distribution are regressed on the Weibull cumulative distribution function. The resulting coefficients are those which minimize the sum of the squares of the differences between the observed and modeled (Weibull) cumulative distributions. USAFETAC uses a nonlinear regression scheme suggested by Heuser, Somerville, and Bean (1980). The initial guess portion of their technique has been improved by a linearization procedure developed by Maj Al Boehm, USAFETAC/DNP, which is summarized below

Let $Q = 1 - P$ (the probability that X is greater than x_T) and substitute this value into Equation (A-7)

$$Q = \exp(-\alpha x_T{}^{\beta}) \qquad \text{(A-9)}$$

Take the natural logarithm of each side of Equation (A-9), which yields

$$\ln Q = -\alpha x_T{}^{\beta} \qquad \text{(A-10)}$$

Equation (A-10) can be rewritten as

$$-\ln Q = \alpha x_T{}^{\beta} \qquad \text{(A-11)}$$

Equation (A-11) is in the form of a regular power function, which can be linearized by taking the natural logarithm of each side.

$$\ln[-\ln Q] = \ln \alpha + \beta \ln x_T \qquad \text{(A-12)}$$

For a power function fit by the method of least squares, the estimates of α and β are obtained by fitting a straight line to the set of ordered pairs ln X_T, ln[-ln Q]. Substituting these values into the normal equations for a straight line, the solution for β becomes

$$\beta = \frac{n\Sigma(\ln x_T)(\ln[-\ln Q]) - (\Sigma \ln x_T)(\Sigma \ln[-\ln Q])}{n\Sigma\ (\ln x_T)^2 - (\Sigma \ln x_T)^2} \qquad \text{(A-13)}$$

The solution for α becomes

$$\alpha = \exp(\overline{\ln[-\ln Q]} - \beta\ \overline{\ln x_T}) \qquad \text{(A-14)}$$

Note that for the linearization technique to be successful, $0 < Q < 1$. All ordered pairs where Q is equal to 1 or 0 should not be used in the regression, because ln[-ln Q] will undefined. It has been found, however, that the simple linear regression minimized the RMS error in ln[ln Q] space and was not necessarily the case when translated to Q space. To approximate minimum errors in Q space, it was necessary to apply a weighting factor ($WF = -Q \ln Q$) to each data point. This weighting factor was later modified to $(Q \ln Q)^2$ by Maj Pershing Hicks, USAFETAC/DNO. This technique provides excellent initial guesses to Heuser, Somerville, and Bean's nonlinear curve fitting procedure.

Variable to be Modeled. Visibility

Function Name. Inverse Linear

General. A second function used by USAFETAC to model visibility is the inverse linear function. This function was first suggested for fitting visibility data by O'Connor of USAFETAC and is described by Friend (1978). The inverse linear function for visibility is given by the equation,

$$P = \frac{1}{F\ x_T + G} \qquad \text{(A-15)}$$

where F and G are modeling coefficients determined from empirical data, x_T is some threshold visibility in meters, and P is the probability that the actual visibility (X) is greater than or equal to x_T.

Inverse Transnormalizing. Values of visibility can be obtained from given probabilities by solving Equation (A-15) for x_T

$$x_T = \frac{1 - G\ P}{F\ P} \qquad \text{(A-16)}$$

Fitting the Curve. The modeling coefficients F and G can be obtained for any observed cumulative distribution of visibility by any good linear regression technique. USAFETAC has used a method described by Forsythe, et al. (1977) quite successfully. The method called DECOMP and SOLVE inverts the normal equations by Gaussian elimination and then solves for the modeling coefficients. The Weibull function has proved itself superior to the inverse linear and is used for USAFETAC's simulation applications.

Variable to be Modeled. Ceiling

Function Name. Burr Curve

General. USAFETAC's basic modeling equation for ceiling is the three-parameter Burr curve. This function was first applied to fitting ceiling data by Bean, Somerville, and Heuser (1979). The Burr curve is given by the equation,

$$P = 1 - [1+(x_T/C)^A]^{-B} \qquad \text{(A-17)}$$

where A, B, and C are modeling coefficients determined from empirical data, x_T is some threshold ceiling in feet, and P is the probability that the actual ceiling (X) is less than or equal to x_T

Inverse Transnormalizing. Values of ceiling can be obtained from given probabilities by solving Equation (A-17) for x_T

$$x_T = (C)[\ (1-P)^{-1/B} - 1\ ]^{1/A} \qquad \text{(A-18)}$$

Fitting the Curve. To obtain the modeling coefficients, A, B, and C, the values of an empirical cumulative distribution are regressed on the theoretical cumulative distribution (Burr curve). The resulting coefficients are those that minimize the sums of the squares of the differences between the observed and modeled (Burr) distribution. USAFETAC uses a nonlinear regression scheme suggested by Heuser, Somerville, and Bean (1980). The initial guess portion of the technique was developed by Capt Emil Berecek, USAFETAC/DNS.

Variable to be Modeled. Ceiling

Function Name. Reverse Weibull Curve

General. USAFETAC uses a slightly different form of the Weibull curve, called the "reverse Weibull," to fit ceiling data than it does for visibility data.

This form was developed by Maj Pershing Hicks, USAFETAC/DNO (Hicks, 1982). The reverse Weibull curve for ceiling data is given by the equation,

$$P = \exp(-\alpha x_T^{\beta}) \qquad \text{(A-19)}$$

where α and β are modeling coefficients determined from empirical distributions, x_T is some threshold ceiling in feet, and P is the probability that the actual ceiling (X) is less than or equal to x_T.

Inverse Transnormalizing. Values of ceiling can be obtained from given probabilities by solving Equation (A-19) for x_T,

$$x_T = [\ \frac{\ln(P)}{-\alpha}\ ]^{1/\beta} \qquad \text{(A-20)}$$

Fitting the Curve. To obtain the modeling coefficients α and β, the weighted linear regression technique developed by Hicks and Boehm of USAFETAC (mentioned in the section on the Weibull model for visibility distributions) is used. Values of α and β are obtained by fitting a straight line to the set of ordered pairs of data ln x_T, ln[-ln P] with the following equations

$$\beta = \frac{\Sigma\ WF\ \Sigma(\ln x_T)(\ln[-\ln P])(WF) - (\Sigma\ WF\ \ln x_T)(\Sigma\ WF\ \ln[-\ln P])}{(\Sigma\ WF\ \Sigma\ WF\ \ln x_T^2) - (\Sigma\ WF\ \ln x_T)^2} \qquad \text{(A-21)}$$

$$\alpha = \exp\ (\ \frac{\Sigma\ WF\ \ln[-\ln P]}{\Sigma\ WF} - \beta\ \frac{\Sigma\ WF\ \ln x_T}{\Sigma\ WF}\ ) \qquad \text{(A-22)}$$

where

$$WF = (P \ln P)^2 \qquad \text{(A-23)}$$

Closeness of fit between the observed and modeled distributions (as measured by RMS difference) for the reverse Weibull curve compares quite favorably to those values achieved by the Burr curve. For very large modeling efforts, a significant amount of computer time can be saved by using the easier and faster linear regression technique for the reverse Weibull function as opposed to the nonlinear method that must be used for fitting the ceiling data to the Burr curve. The computer time saved translates to lower model development costs. In certain applications, the lower costs are more important than the slightly better accuracy that can be achieved using the Burr curve.

Variable to be Modeled. Ceiling

Function Name. Log-Cubic Equation

General. A second function that can be used to model cumulative distributions of ceiling is the four-parameter log-cubic equation. This function was first adapted for use in fitting ceiling data by O'Connor of USAFETAC and is described by Friend (1978). The log-cubic equation for ceiling is given by

$$P = C_1 + C_2 (\ln x_T) + C_3 (\ln x_T)^2 + C_4 (\ln x_T)^3 \qquad \text{(A-24)}$$

where C_1 - C_4 are modeling coefficients determined from empirical data, x_T is some threshold ceiling in feet and P is the probability that the actual ceiling (X) is greater than or equal to x_T.

Inverse Transnormalizing. The log-cubic equation is of limited use for simulation applications for two important reasons. First, Equation (A-19) cannot be uniquely solved for x_T from given values of P. It is entirely possible that three very realistic solutions of x_T could exist for a particular value of P. Second, the values of P are not bounded between 0 and 1 with this polynomial. This unboundedness is an undesirable property because it allows the function to return probabilites that are negative or greater than 1.

Fitting the Curve. To obtain the modeling coefficients C_1 - C_4, linear regression techniques must be used. USAFETAC has used two methods successfully (Forsythe, et al., 1977). One method is a singular value decomposition scheme called SVD. The second method, called DECOMP and SOLVE, inverts the normal equations by Gaussian elimination and then solves for the modeling coefficients.

Appendix B

ROOT-MEAN-SQUARE (RMS) DIFFERENCE AS A MEASURE OF CLOSENESS OF FIT

A major part of USAFETAC's simulation effort involves fitting the observed probability distributions of meteorological variables to mathematical functions. This procedure is referred to as distribution fitting, curve fitting or modeling. The resulting fitted mathematical functions then reside within the main environmental simulat:on model. Some measure of the accuracy or "closeness" of these fits is needed in order to be able to make inferences from the results of the environmental simulation model.

When data are expressed as values on a continuous scale, closeness of fit between observed distributions and the modeled (mathematical function) distributions can be expressed as a RMS value. RMS difference is calculated by the following equation:

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (O_j - T_j)^2} \qquad \text{(B-1)}$$

where O_j and T_j are individual elements of the observed and theoretical distributions, and N is the total number of data pairs. For example, the following cumulative frequency distribution for visibility at Leipheim, Germany, February, 0000 LST was fitted to the Weibull distribution

Threshold Visibility v_T (SM)	Observed Cumulative Frequency Vsby $\leq v_T$ (%)	Weibull Cumulative Frequency Vsby $\leq v_T$ (%)	Residual (%)	Residual Squared
0.0	0.0	0.0	0.0	0.00
0.5	12.4	11.9	0.5	0.25
1.0	24.3	21.8	2.5	6.25
2.0	35.2	37.9	-2.7	7.29
3.0	50.6	50.5	0.1	0.01
4.0	64.2	60.3	3.9	15.21
5.0	66.2	68.1	-1.9	3.61
6.0	76.1	74.3	1.8	3.24

The sum of the residuals squared is 35.86, and the total number of data pairs is eight. Using these values in Equation (B-1) yields

$$\text{RMS} = \sqrt{35.86/8}$$

$$= \sqrt{4.4825}$$

$$\text{RMS} = 2.1$$

One can normally surmise that the lower the RMS value the better the fit, as illustrated by Figures 12 and 13. These figures were already discussed in Chapter 5. Figure 13 compared the observed and modeled relative frequency distributions of sky cover for Moscow, RS, November, 0600 LST. The RMS value was very low (0.5), and the modeled curve duplicated the observed distribution quite well. Figure 12 compared the observed and modeled relative frequency distributions of sky cover for Chiganak, RS, November, 0900 LST. The RMS value was 5.3. Although the fit was able to capture the general shape of the distribution, some very large errors did result.

There is a disadvantage in using RMS, however. A modeled distribution might have a relatively low RMS but not adequately reproduce the shape of the observed distribution as illustrated by Figure B-1.

Figure B-1 compares the observed and modeled distributions of sky cover at Feddosiya, RS, 1800 LST, August. The RMS for this fit was 2.8 which is not altogether bad for this type of distribution, but note that the modeled curve has not adequately reproduced the relative maxima in the 2/8 and 6/8 coverage categories. The RMS value can be quite helpful in determining closeness of fit but it is not as effective as a visual comparison of the two distributions.

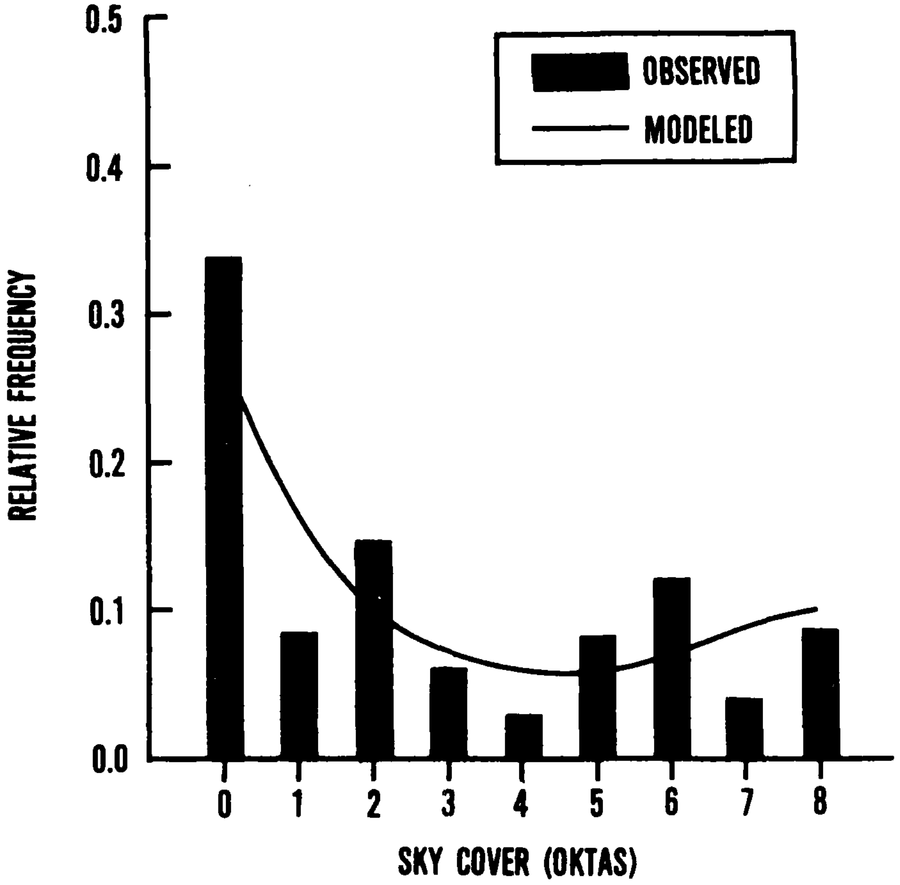


Figure B-1. Relative Frequency Distribution of Cloud Cover at Feddosiya, RS, August, 1800 LST. The observed distribution and the Johnson S_B curve fit to that distribution are shown. The RMS between the observed distribution and modeled CDF is 2.8 percent, and the maximum difference is 4.5 percent.

Appendix C

SERIAL CORRELATION IN FIRST-ORDER MARKOV MODELS

A defining property of the Ornstein-Uhlenbeck process (Parzen, 1962) is that

$$\mathrm{Cov}[x_{T+\Delta t}, x_T] = \rho\sigma_{x_{T+\Delta t}}\sigma_{x_T} = \alpha e^{-\beta\Delta t} \qquad \text{(C-1)}$$

which equals ρ for x distributed N(0,1). Hence, for N(0,1) x,

$$\rho = \alpha e^{-\beta\Delta t} \qquad \text{(C-2)}$$

Applying the boundary condition that $\rho = 1$ when $\Delta t = 0$ gives the result $\alpha = 1$. So,

$$\rho = e^{-\beta\Delta t} \qquad \text{(C-3)}$$

If ρ_1 is defined as the correlation at unit time, where $\Delta t = 1$, then

$$\rho_1 = e^{-\beta} = \text{const} \qquad \text{(C-4)}$$

Consider ρ at n time steps, $n\Delta t$

$$\rho = e^{-\beta n\Delta t} = (e^{-\beta})^{n\Delta t} = \rho_1{}^{n\Delta t} \qquad \text{(C-5)}$$

which is the characteristic Markovian correlation decay equation.

Appendix D

THE PARTIAL AUTOCORRELATION FUNCTION

Let the autocorrelation function of a stochastic process be denoted by ρ_k for lag k.

The partial autocorrelation, often called the partial autocorrelation function, is a function of the autocorrelation ρ_k. Let ϕ_{kj} be the jth coefficient in an autoregressive (AR) process of order k, so that ϕ_{kk} is the last coefficient in the series. The AR process must satisfy the equation,

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k} \qquad j = 1,2, \ldots k \tag{D-1}$$

which leads to the Yule-Walker equations,

$$\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{vmatrix} \quad \begin{Vmatrix} \phi_{k1} \\ \phi_{k2} \\ \cdots \\ \phi_{kk} \end{Vmatrix} = \begin{Vmatrix} \rho_1 \\ \rho_2 \\ \cdots \\ \rho_k \end{Vmatrix} \tag{D-2}$$

or

$$\underline{P}_k \, \boldsymbol{\Phi}_k = \underline{\rho}_k \tag{D-3}$$

Solving these equations for k = 1, 2, 3..., successively, gives the result,

$$\phi_{11} = \rho_1 \tag{D-4}$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{D-5}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} \tag{D-6}$$

where $|\cdot|$ represents the determinant. For ϕ_{kk}, the determinant in the numerator has the same elements as the determinant in the denominator, but with the last column replaced by ρ_k.

The quantity ϕ_{kk}, which is a function of the lag k, is called the *partial autocorrelation function*. In an AR(p) process, i.e., an autoregressive process of order p, the partial autocorrelation function ϕ_{kk} will be non-zero for $k > p$. This is another way of saying that the partial autocorrelation function ϕ_{kk} of a pth order AR process will exhibit a cutoff after lag $k = p$.

The partial autocorrelation functions ϕ_{kk} described above are *theoretical*. In practice, one must *estimate* these theoretical autocorrelations from sample data using the methods of Box and Jenkins (1976). The estimated or *sample* partial autocorrelations are denoted by $\hat{\phi}_{kk}$.

LIST OF ABBREVIATIONS AND ACRONYMS

AAF	Army Air Field
Abv or ABV	Above
AFB	Air Force Base
AMS	American Meteorological Society, a professional society for meteorologists
AR	Autoregressive, a type of stochastic process model
ARMA	Autoregressive-moving average, a type of stochastic process model that includes both autoregressive (AR) and moving average (MA) terms
Avg or AVG	Average
AWS	Air Weather Service, a technical service of the Air Force's Military Airlift Command
BLM	Boundary Layer Model, a numerical weather prediction model
Blw or BLW	Below
COLOSSUS	An informal name for the Military Airlift Command's M-14 airlift operations simulation
const	Constant
Cor Coef	Correlation coefficient
cos	Trigonometric cosine function
Cov	Covariance
CPU	Central processing unit, that part of a computer which accomplishes arithmetic and logical operations
DCFLOS	Dynamic cloud-free line-of-sight, a simulation model designed to calculate the probability of having a cloud-free line-of-sight between two moving points for a duration of time
DECOMP	Gaussian decomposition subroutine, used with SOLVE for Gaussian elimination solutions
e	The mathematical e, the base of the system of natural logarithms, having the approximate value 2.71828
E[·]	Expectation operator
END	Equivalent normal deviate, the standard normal variable
exp(·)	Exponential operator; $\exp(a) = e^{a}$
ft	Feet, a unit of linear measure
GCD	Great circle distance
GMT	Greenwich Mean Time
INT(·)	Integer operator, indicates the largest integer less than or equal to the value stated in the argument
JSKY1	Joint sky cover probability model for the lag problem, developed by USAFETAC and described in this report
JSKY2	Joint sky cover probability model for the spatial problem, developed by USAFETAC and described in this report
km	Kilometers, a unit of linear measure
LFM	Limited Area Fine Mesh Model, a numerical weather prediction model
log	Common logarithm, base 10
ln	Natural logarithm, base e
LST	Local Standard Time
LUSQRT	Cholesky reduction or "square root" method for factoring a real, symmetric, positive definite matrix into a lower triangular matrix and its transpose
M-14	A simulation of the total military airlift system in peace and war; also informally called COLOSSUS; developed by the Military Airlift Command

MA	Moving average, a type of stochastic process model
MAC	Military Airlift Command, a specified command of the US Air Force
MORS	Military Operations Research Society; or Military Operations Research Symposium, the latter conducted by the former
MULTRI	Multivariate triangular matrix environmental simulation model developed by USAFETAC and described in this report
NM	Nautical miles, a unit of linear measure
N(0,1)	Normal distribution with mean of zero and variance of unity
Obs	Observation or observations
OL-A	Operating Location-A, a type of Air Force unit
Pct or PCT	Percent
PPM	Pearson product moment formula for calculation of the correlation coefficient from sample data
$\Pr(\cdot)$ or $\Pr\{\cdot\}$	Probability operator, representing the probability of the condition in parentheses or braces
RANDCV	Routine for generation of a correlated vector of random normal numbers
RMS	Root mean square
RS	Russia
RUSSWO	Revised Uniform Summary of Surface Weather Observations, a statistical tabulation of historical weather data for a single location, produced by USAFETAC
S_B	Single bounded, a member of Johnson's family of curves
sin	Trigonometric sine function
SD	Scale distance, especially in regard to Gringorten's Model-B
SM	Statute miles, a unit of linear measure
SOLVE	Back substitution routine, used with Gaussian decomposition routine DECOMP for Gaussian elimination solutions
STRICOM	Strike Command, a unified command of the United States
SVD	Singular value decomposition, an alternative to linear regression for determining coefficients of a linear equation
tan	Trigonometric tangent function
tanh	Trigonometric hyperbolic tangent function
USAFETAC	United States Air Force Environmental Technical Applications Center, the applied climatological arm of the Military Airlift Command's Air Weather Service
V1S1	Single-variable, single-station environmental simulation model developed by USAFETAC and described in this report
V2S1	Two-variable, single-station environmental simulation model developed by USAFETAC and described in this report
WMO	World Meteorological Organization
2DFLD	Two-dimensional field simulation model, an environmental simulation model developed by USAFETAC and described in this report
$\cup$	Union, as in $A \cup B$, interpreted as "A or B"
$\cap$	Intersection, as in $A \cap B$, interpreted as "A and B"
ε	Member of (set notation)
Σ	Summation operator
$\int$	Integral operator
$<$	Less than
$>$	Greater than
$(A \mid B)$	Condition A, given condition B; used in expressing conditional probabilities
$(\cdot,\cdot)$	Open interval; does not include the end points
$[\cdot,\cdot]$	Closed interval; does include the end points
$\lvert\cdot\rvert$	Determinant operator
$\hat{}$	Estimate indicator; $\hat{Y}$ is an estimate of the value of Y
π	Geometric pi, the ratio of the circumference of a circle to its diameter

FILM

2-83